University of Cape Town

School of Management Studies

How Does Frame-of-Reference Training Increase Rater Accuracy? A Test of Potential Explanatory Mechanisms.

Natasha Baret

(BRTNAT007)

A dissertation submitted in partial fulfilment of the requirements for the award of the Degree of Master of Social Science in Organisational Psychology

Supervisor: Francois de Kock

October 2018

COMPULSORY DECLARATION:

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works of other people has been attributed, cited and referenced.

Signature: Signed by candidate **Date:** 20 October 2018

Acknowledgements

There are a number of individuals who contributed to the completion of this research dissertation. I would firstly like to express my sincere gratitude towards my supervisor, Francois de Kock, for his invaluable insight, expertise and support throughout this process. Thank you for always guiding me in the right direction.

I would also like to extend my thanks to Filip Lievens (Singapore Management University) who was kind enough to share his training protocol which provided the basis for my training material. Your expertise in this field is duly noted.

Thank you to Alexandra Marsh and Sandhia Raghubeer for being kind enough to assist me with my data collection. Without you, I would not have been able to collect my data in the time constraints that are imposed in the Master's framework.

Most importantly, I would like to thank my parents, Ann and Christian, as well as my brother, Christophe, who have supported me both financially and personally throughout my tertiary education. You have always encouraged me to strive for great heights and to achieve to the best of my ability. I would not be where I am today without my family by my side.

Abstract

In the job interview literature, the positive effect of Frame-of-Reference (FOR) training on interviewer rating accuracy is well documented. However, *how* this training method increases rating accuracy is not well understood. The present study focused on rater individual difference characteristics as potential explanatory mechanisms for the effects of FOR training on accuracy. The researcher hypothesised that FOR training would enhance raters' dispositional reasoning, rating motivation and self-efficacy, which, in turn, would increase their rating accuracy. A post-test only experimental research design was used in a sample of 32 students from a South African university. Participants were randomly assigned to the FOR training intervention or the no-training condition. Participants were required to rate three videotaped candidates on an interview competency and completed various individual difference measures. The FOR training intervention positively affected rating accuracy and findings suggest this effect may occur because of the influence of FOR training on dispositional reasoning, rather than FOR training enhancing rater motivation or self-efficacy. Study limitations and recommendations for future research are noted.

Key words: rating accuracy, realistic accuracy model, frame-of reference training, dispositional reasoning, rater motivation, rater self-efficacy.

Table of Content

List of Tables

List of Figures

Introduction

The employment interview is ubiquitous in the personnel assessment and selection domain (Huffcutt, Culbertson, & Weyhrauch, 2011; Levashina, Hartwell, Morgeson, & Campion, 2014; Macan, 2009). Interviews require interviewers[1] to observe and rate applicants on their predicted job performances based on their personality traits, work experience and work behaviours (Macan, 2009). Thereafter, these ratings are used for selection, development and termination decisions (Bernardin & Villanova, 2005; Jiang, Lepak, Hu, & Baer, 2012). As such, the accuracy of these ratings has important implications for the effectiveness of organisations (Christiansen, Wolcott-Burnam, Janovics, Burns & Quirk, 2005).

As rater accuracy is of importance for organisational effectiveness, it is not surprising that research has largely focused on the development of training approaches to improve accuracy (Roch, Woehr, Mishra, & Kieszczynska, 2012). One such training approach is frame-of-reference (FOR) training (Bernardin & Buckley, 1981). This training typically involves participants observing job performance behaviours in a videotaped recording, after which they are then asked to rate the performance (Roch et al., 2012). The trainer then informs the participants of the correct ratings and the rationale behind each rating (Roch et al., 2012).

The effectiveness of FOR training's to increase rater accuracy remains undisputed, as there has been an abundance of research supporting FOR training and its successes (Roch et al., 2012; Woehr & Huffcut, 1994). FOR training has been argued to increase rater accuracy by imposing correct performance schemas on raters (Lievens, 2001). The imposed schemas

[1] In this paper, we use the term 'judges', 'raters', 'assessors' and 'interviewers' interchangeably as well as the term 'ratee', 'candidate', 'target' and 'interviewees' interchangeably

provide raters with a common frame-of-reference, which is intended to get all raters on the same metric when assigning ratings (Uggerslev & Sulsky, 2008).

Limited empirical research has been conducted to investigate how FOR training manages to increase rater accuracy. One line of research has focused on the training methodology (Hauenstein & McCusker, 2017) and found that accuracy increased through the repeated practice and feedback sessions, which is typically involved in the FOR training process (Bernardin & Buckley, 1981). Another line of research (e.g., De Kock, Lievens, & Born, 2018; Powell & Bourdage, 2016; Powell & Goffin, 2009) has focused on individual difference characteristics that may enhance behavioural cue detection and cue utilisation in the rating context. According to Funder's Realistic Accuracy Model (RAM; 1995; 1999; 2012) the ability to correctly detect and utilise cues is fundamental in the judgement accuracy process.

One such individual characteristic is dispositional reasoning, defined as a complex knowledge of traits, behaviours and the potential of situations to elicit traits into behaviours (De Kock et al., 2018). Dispositional reasoning is a relatively strong predictor of rater accuracy (Christiansen et al., 2005; De Kock, Lievens, & Born, 2015). As such, FOR training may improve accuracy through its potential effects on dispositional reasoning (Powell & Bourdage, 2016; Powell & Goffin, 2009). Although empirical studies (Powell & Bourdage, 2016; Powell & Goffin, 2009) have supported the notion that FOR training increased rating accuracy, and dispositional reasoning was associated with higher accuracy scores (Powell & Bourdage, 2016), these studies have shown no significant difference in dispositional reasoning scores between those who were trained and those who were not (Powell & Bourdage, 2016; Powell & Goffin, 2009). These results indicated that FOR training therefore did not improve accuracy through dispositional reasoning. If the strongest predictor of rater accuracy does not influence the effectiveness of FOR training in enhancing rater accuracy,

then the question remains: Which other individual characteristics may account for how FOR training increases rater accuracy?

The present study explores other individual characteristics that may explain how FOR training increases accuracy. For example, training raters has the potential to increase their motivation to rating accurately (Harris, 1994). Increased motivation would increase their attention in detecting behavioural cues during the rating process and consequentially may affect the accuracy of their ratings (Funder, 1999). Furthermore, for individuals to rate accurately, they need to have the confidence and a sense of personal mastery to correctly utilise the skills and knowledge gained during the training (Wood & Marshall, 2008). This confidence is referred to as rater self-efficacy (Bandura, 1986; Bernardin & Buckley, 1981). Studies of both rater motivation and rater self-efficacy may offer insight into how FOR training increases rater accuracy. However, these individual differences remain relatively unexplored in contemporary research despite their promising findings and theoretical underpinnings (De Kock et al., 2018).

This investigation intends to deepen understanding of how FOR training may increase rater accuracy by investigating the influence of rater individual characteristics. Firstly, the study aims to replicate previous findings to confirm whether FOR training will influence dispositional reasoning as well as rating accuracy. Previous findings (Powell & Bourdage, 2016; Powell & Goffin, 2009) that show limited effects of FOR training on dispositional reasoning are counterintuitive to the important role of this construct in rater accuracy (Christiansen et al., 2005; De Kock et al., 2015). Secondly, the study will determine the influence of rater motivation and rater self-efficacy on rater accuracy during the FOR training, as both constructs lack empirical investigation. The research question is, therefore, as follows: *What are the individual difference variables that may explain the effect of FOR training on rater accuracy?*

Literature Review

The literature review will firstly present the theoretical framework used to understand rater accuracy, namely Funder's Rater Accuracy Model (RAM; 1995; 1999; 2012). Following this, the training approaches used to develop raters are discussed. Particular attention is paid to the FOR training approach. Next, the existing literature on the variables being investigated in this study are discussed, namely rater dispositional reasoning, motivation and self-efficacy, and their effect on accuracy during the FOR training. A graphical representation of the conceptual framework for this study is presented following the literature review (See Figure 3).

Rater Accuracy

Rater accuracy refers to the relationship between what is perceived by the rater and what is in fact reality (Funder, 1999). In a practical context, it refers to the overlap between the "true score" ratings made by subject matter experts and those made by raters in practice (Engelhard, 1996). For the purpose of this study, rater accuracy is defined as the extent to which interviewers reflect the "true score" in their assigned ratings.

As previously mentioned, rater accuracy is crucial to the quality of decision that affect both the individuals being rated and the organisation (Funder, 1999; Schmitt & Chan, 1998). These ratings have important implications for selection and promotion decisions. Through accurate ratings of candidates, organisations are able to select candidates that will be most successful and effective in their assigned roles (Funder, 1999).

The implications of quality ratings on individual and organisation decisions have led many researchers to explore this phenomenon. Prior to the early 1990s, researchers had largely ignored the accuracy of ratings and had focused rather on biases and errors in judgement and ratings (Funder, 1999). Rater accuracy research was revived in the late 1990s following Funder and West's (1993) seminal work. They called for researchers to refocus their efforts to accuracy in judgements, as at the time of their publication there was a diverse

range of focus and a lack of consensus pertaining to factors influencing the quality of ratings. Following this, research on rater accuracy rapidly increased (Bernardin, Tyler, & Villanova, 2009; Christiansen et al., 2005; De Kock et al., 2015; De Kock et al., 2018; Engelhard, 1996; Funder, 1995; Funder, 1999; Funder; 2012; London, Mone, & Scott, 2004; Mero & Motowidlo, 2003; Powell & Bourdage, 2016; Powell & Goffin, 2009; Sulsky & Day, 1994).

It is important to note that these two research streams, namely rating error and rater accuracy, focus on different research questions. Research that focuses on rater errors and biases aims to investigate whether the rating process followed normative rules based on policies, formal logic, mathematics and statistics (Funder, 2012). Simplified, rater error research attends to whether the rating process was correct based on logic and theory. In contrast, research that focuses on rater accuracy aims to investigate whether the ratings and judgements made were correct (Funder, 2012).

Following the revival of rater accuracy research, a key research area has focused on individual differences amongst raters that affect their accuracy (De Kock et al., 2018; Graves, 1993). It is still unclear which individual differences facilitate rater accuracy (Guion & Highhouse, 2011). What is clear is that there are good judges that make accurate ratings of targets, and there are poor judges that make inaccurate ratings of targets (De Kock et al., 2018). The theoretical underpinnings of a 'good' judge rests in Funder's Realistic Accuracy Model.

Realistic Accuracy Model. One of the most insightful contemporary theoretical framework models of rater accuracy is Funder's Realistic Accuracy Model (RAM; Funder, 1995; 1999; 2012; See Figure 1). Within RAM, it is argued that accurate ratings rely on a judge's ability to detect and utilise behavioural cues, and that there are four stages in the judgement process that need to be met in order for accurate judges to occur (see Figure 1; Funder, 2012).

In the first stage, *relevance*, the ratee is required to emit a behaviour that is relevant to a trait being judged in the rating context (Funder, 1995). For example, in order for a rater to judge whether a ratee is friendly, the ratee needs to emit a friendly behaviour such as smiling or laughing.

In the second stage, *availability*, the rater requires the behaviour to be available to them (Funder, 1995). Using the same example as above, the friendly ratee needs to demonstrate the friendly behaviour in the rating context. If the behaviour, such as smiling, is exhibited outside of the rating context, the behaviour is not available to the rater. In turn, the rater is not able to make use of this behavioural information and will not be able to judge the ratee as being friendly (Funder, 2012).

In the third stage, *cue detection*, the rater should detect the behaviour (Funder, 1995). For example, the rater would need to detect the ratee smiling, laughing or any friendly demeanour. If the rater is distracted, impartial or unperceptive then accurate judgements will be hindered, as the rater is not been able to detect the behaviour necessary for accurate judgements (Funder, 2012).

In the fourth and final stage, *cue utilisation*, the judge should correctly utilise the behaviour being demonstrated to make an accurate judgment of the ratee (Funder, 1995). For example, the rater would need to utilise the behavioural cue of smiling and correctly interpret it as being friendly and not misinterpret it as the candidate being insincere (Funder, 2012).

Funder (1995; 1999; 2012) stated that in order for accurate judgements to occur, all four stages need to be met otherwise an inaccurate judgement will occur. To further demonstrate, a candidate being considered for a managerial position would need first to display a behaviour that is relevant to a managerial trait being judged, such as effective communication. This behaviour could be high verbal skills. Second, the rater would need the high verbal skills to be made available to them during the interview. The candidate would

need to communicate in the interview in a confident and clear manner, and not mumble. Lastly, the rater would need to detect the high verbal skills being exhibited and utilise the information to judge accurately whether the candidate communicates effectively.

According to the RAM model there are various moderators that will influence the judgement process described above (De Kock et al., 2015; Funder, 2012). Firstly, a *good target* is needed to display behaviour that is a true reflection of their personality. Secondly, a *good trait* will be more perceptible and easier to detect. Thirdly, *good information* is needed to allow the rater to observe the target in different contexts over a period of time. Lastly, a *good judge* will be able to make accurate judgements and decisions in a short time frame. In addition, a *good judge*'s behaviour, such as constant eye contact and expressed warmth and sympathy, will affect the availability and relevance of cue needed (Funder, 2012). The judge's behaviours will create situations that may provoke more relevant personality cues from targets, which would further facilitate the accurate judgement process (Letzring, 2008; Lievens, Schollaert, & Keen, 2015).

Funder's (1995; 1999; 2012) RAM framework is useful in understanding the judgement process involved in rater accuracy. Any attempts to improve accuracy needs to take into account the four stages described, otherwise their attempts will be futile (Funder, 1995). Figure 1 below illustrates the judgement process described as well as the moderators. The next section of the literature review discusses training approaches in their attempts to improve the accuracy of raters.

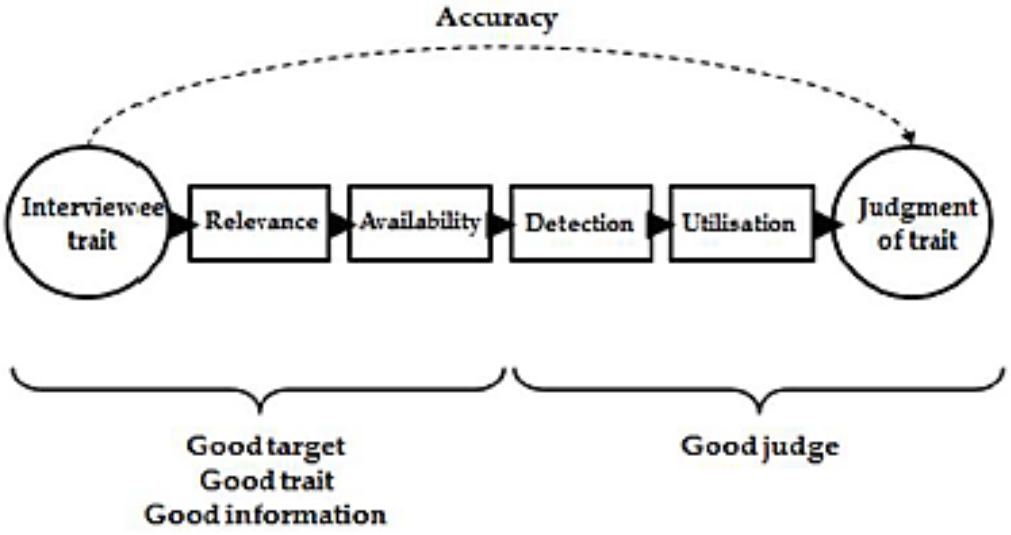

Figure 1. The Realistic Accuracy Model: Processes and moderators (Funder, 2012).

Approaches to Rater Training

The notion of training raters has been popular amongst researchers for several decades, starting with Driver (1942) who, without empirical evidence, postulated that training raters is critical to the rating process. The reason for the abundance of research in this area is that researchers have consistently questioned the quality of ratings due to the subjectivity of raters (Hauenstein & McCusker, 2017). As a result, research into the design of training has been given a considerable amount of attention (Woehr & Huffcut, 1994). Three rater training programmes in particular have received the most attention, namely: rater error training, performance dimension training, and frame-of-reference training.

Rater error training. A review of rater training research reveals that early approaches to rater training focused on reducing or eliminating rater errors. This was achieved by making raters aware of errors and heuristics, such as leniency, central tendency and halo effect, and then training the raters to avoid the errors in the rating process (Athey & McIntyre, 1987; Hedge & Kavanagh, 1988; Latham, Wexley, & Pursell, 1975; McIntyre, Smith, & Hassett, 1984).

Rater error training (RET) is an indirect approach to improving accuracy, as RET addresses the errors rather than increasing accuracy directly. Whilst RET was effective in training raters to avoid these errors (Woehr & Huffcut, 1994), studies showed that RET may have reduced rater accuracy as errors used by raters may have in fact resulted in accurate judgements of the candidate's performance (Bernardin & Buckley, 1981; Bernardin & Pence, 1980; Gigerenzer, 2008). Simplified, there are instances when errors and heuristics used by raters allow them to make accurate ratings, therefore by reducing or eliminating these errors, RET reduces accuracy (Gigerenzer, 2008).

Performance dimension training. Performance dimension training (PDT) focuses on the performance dimensions that raters target in their judgements (Lievens, 1998; Smith, 1986; Woehr & Huffcut, 1994). The premise of this training approach rests in the notion that raters form judgements on behaviour that is observed in the moment, as opposed to a later stage when the evaluation may be required (Hastie & Park, 1986; Woehr & Feldman, 1993). PDT, therefore, focuses on training raters to make judgements in the present moment by familiarising the raters with what the performance dimension being targeted is and providing them with specific definitions of what the performance dimension is (Woehr & Huffcut, 1994). By doing this, PDT makes information available, from which raters are able to use when making judgements in the moment. According to a meta-analysis conducted by Woehr and Huffcut (1994), whilst PDT has received considerable amount of attention from researchers and practitioners in regard to rater training, on average PDT only leads to a slight increase in accuracy (d = .13) compared to other training methods.

Evaluation of earlier training approaches. Whilst both RET and PDT focus on developing raters, these two training approaches are less favoured compared to a third training approach named frame-of-reference (FOR) training (Woehr & Huffcut, 1994; Roch et al., 2012). A reason for this favourability is that RET and PDT do not consider the

schematic framework that raters use when making judgement decisions, whereas FOR training does (Woehr & Huffcut, 1994). FOR training focuses on the schemas raters use and provides the raters with accurate and expert schemas which they can use to make accurate ratings. Confirming this claim, Hedge and Kavanagh (1998) compared the effect of RET training and FOR training, and found that the FOR training increased accuracy more so than RET training. Due to this favourability and empirical support of FOR, which will be discussed shortly, the present study will focus its attention on the FOR training approach.

Frame-of-Reference Training

FOR training, introduced by Bernardin and Buckley (1981), is the most favoured training method to increase rater accuracy by influencing the way raters encode, organize and recall information during the rating process (Roch et al., 2012; Sulsky & Day, 1994; Woehr & Huffcutt, 1994). In a meta-analysis on rater training conducted by Roch et al. (2012), the researchers noted that considerably more empirical studies investigated FOR training methods compared to RET or PDT methods. They also claimed that FOR training is particularly effective in all functions that rely on ratings, such as performance appraisals, assessment centres, selection test cut scores, employment applications, competency modelling, job analysis and employment interviews (Roch et al., 2012).

The goal of FOR training is to train raters to use a common and informed frame-of-reference of the performance dimension being assessed (Athey & McIntyre, 1987; Woehr, 1994). To achieve this common frame-of-reference, FOR typically consists of the trainer first identifying job performance dimensions and their corresponding evaluative standard. The trainer then provides examples of these performance dimensions and corresponding evaluative standards based on good, average and poor behaviours (Roch et al., 2012; Woehr & Huffcutt, 1994). Following this, raters assess and rate certain performance dimensions of a target, typically portrayed in a videotape. Upon completion of the assessment and ratings, the

trainer then informs the raters of the correct ratings and the rationale behind each rating, allowing the rater to gain more accurate knowledge regarding the behaviours through practice and feedback (Roch et al., 2012). Holistically, the FOR training process will provide the raters with a common frame-of-reference of a specific performance dimension, which they will use in the rating contexts going forward.

Effect on rating outcome. The conceptual understanding of how FOR increases accuracy is that (a) FOR assists raters in understanding which behaviours reflect a certain level of performance on specific job dimensions (Roch et al., 2012) and (b) FOR creates performance prototypes, or samples, that allow raters to accurately categorize the target's performance when assigning ratings (Hauenstein & Foti, 1989; Ilgen, Barnes-Farrell, & McKellin, 1993; Sulsky & Day, 1994; Woehr, 1994).

Moreover, the FOR training approach enforces new and more accurate schemas on raters, thereby correcting possible incorrect schema-based processes used in making judgements (Lievens, 2001). These new and more accurate schemas will then be used in the cue detection and cue utilisation processes according to RAM (Funder, 1995; 1999; 2012), thereby enhancing the judgement process involved in rating accuracy. Research has consistently confirmed that the FOR training process has led to an increase in rating accuracy (Roch et al., 2012; Woehr & Huffcut, 1994).

Empirical findings. A seminal study that supported the effect of FOR training on rater accuracy was conducted by Woehr and Huffcut (1994). They provided an integrated review of rater training literature at the time of their publication by conducting a meta-analysis. They investigated existing studies on the various different rater training approaches and reported that FOR training led to the largest overall increase in rater accuracy and held a considerably large effect size (d = .83; Woehr & Huffcut, 1994). At the time of their

publication, they noted that their study was limited by the small quantity of studies included in their meta-analysis. This potentially affected their study's results.

Following this limitation, a follow up meta-analysis was conducted by Roch et al. (2012). They aimed to expand further on the original meta-analysis by including more empirical studies than Woehr and Huffcut (1994) and to investigate the current state of FOR training at the time of their publication. The former meta-analysis included 29 manuscripts surrounding rater training. Roch et al. (2012) included 90 manuscripts, of which 57 focused on FOR training. Roch et al. (2012) reported a more modest average effect size of d = .50. Whilst their effect size was smaller than the one reported by Woehr and Huffcut (1994), Roch et al. (2012) argued that it was a more realistic portrayal of the effect FOR has on rater accuracy. Regardless, the effect size of .50 is still considered to be of moderate size, which suggests positive expectations of FOR increasing rater accuracy.

Another key empirical finding which provides support towards the effectiveness of FOR in increasing the accuracy of ratings, was conducted by Lievens (2001), who investigated two different types of rater training strategies (N = 390). The study compared a data-driven rater training strategy, aimed at improving the rating behaviour of the raters observing the targets, and the FOR schema-driven assessor training, aimed at improving the schemas and frame-of-reference of the raters. Lievens (2001) reported that those who received the FOR schema-driven training approach achieved higher accuracy scores than those who received a data-driven training approach and no-training. In addition, the study included both students (n = 229) and managers (n = 161), and their findings regarding FOR training resulting in higher accuracy scores applied to both students and managers used in their sample (Lievens, 2001). This implies that FOR training may have an effect on rater accuracy in both a laboratory context as well as in a practical context, such as performance appraisal and employment interviews (Lievens, 2001).

A more recent study supporting the effect of FOR training was conducted by Powell and Bourdage (2016). The authors reported that following FOR training, participants (N = 144) scored higher accuracy scores than those who had no training. Interestingly, the authors focused their FOR training to target two RAM (Funder, 1995) framework stages specifically, namely cue detection and cue utilisation, in order to increase accuracy amongst participants. The authors reported that participants in the cue utilisation FOR training condition scored higher accuracy scores (mean accuracy r = .39) than participants in the cue detection FOR training condition (r = .27; Powell & Bourdage, 2016). This suggests that FOR training is more effective in increasing raters' ability to utilise cues in the judgement process involved in rater accuracy Based on the above mentioned empirical findings, the following hypothesis is posited:

> **Hypothesis 1 (H1):** Participants who receive FOR training will have higher rating accuracy scores than ratings of participants who receive no training.

Rater Individual Differences in Frame-of-Reference Training

As previously mentioned, research indicates that there are particular individual characteristics that influence a rater's ability to accurately judge a target's behaviour and personality (Christiansen et al., 2005; De Kock et al., 2018). Considering that individual characteristics influence the rating process, it is credible to assume that these individual characteristics would influence the effectiveness of FOR training in enhancing accuracy.

The purpose of this study was to investigate three individual characteristics that have been argued either to be the highest predictor of rater accuracy, namely dispositional reasoning (Christiansen et al, 2005; De Kock et al., 2015), or characteristics that have promising theoretical foundations however lack contemporary empirical research, namely rater motivation and rater self-efficacy (De Kock et al., 2018).

Dispositional reasoning. Dispositional reasoning is defined as an individual's understanding of how personality traits, behaviours and situations manifest into observable behaviours (De Kock et al., 2015). Dispositional reasoning has been shown to differentiate between inaccurate raters and accurate raters (Christiansen et al., 2005; De Kock et al., 2015). A good rater uses dispositional reasoning to process observable behaviours into accurate trait inferences, thereby resulting in accurate ratings and judgments (De Kock et al., 2015).

There are three sub-components within the dispositional reasoning construct, namely: *trait induction* (the ability to comprehend how personality traits underlie behaviours); *trait extrapolation* (the ability to comprehend how traits and their behaviours naturally co-vary); and *trait contextualisation* (the ability to comprehend how situations and traits relate to and influence one another) (De Kock et al., 2015; See Figure 2).

Trait induction. Trait induction refers to a judge's ability to comprehend how personality traits are manifested in behaviours, and thus a good judge would be able to comprehend the links between the observable behaviour and traits (De Kock et al., 2015). An example would be for the judge to understand correctly that someone who is shy and does not talk much in a social context, will most likely be an introvert (Goldberg, 1992).

The theoretical foundation of trait induction is trait theory. Trait theory proposes that an individual's traits are habitual patterns that influence behaviour and are stable over time (Allport, 1961; Esyenck, 1970). The argument of how trait induction increases rater accuracy rests in the belief that when an individual attempts to compile inferences of a target being rated, the individual evaluates the target's behaviour according to trait categories (Kihlstrom & Hastie, 1997) and the behaviour is then integrated with situational information (Trope, 1986).

If a rater is able to perform behaviour-trait inferences correctly, then they would be able to compile an accurate overall impression of the ratee (De Kock, et al., 2015). Figure 2

depicts the path link between behaviour and traits found in trait induction. An empirical study (see De Kock et al., 2015) revealed that trait induction predicts rater accuracy, however only to a slight effect (.14).

Trait extrapolation. Trait extrapolation refers to the understanding of how traits and their behavioural manifestations naturally co-vary (De Kock et al., 2015). This ability would allow a rater to observe a target's behaviours and underlying traits and thereafter obtain a wider judgement by filling in gaps of information using their understanding of trait co-variation (see Figure 2). For example, if an accurate rater has thirty minutes to judge a candidate and the candidate portrays the trait of being honest in the interview, the rater may assume the candidate will also be reliable without observing the trait, as the two traits co-vary (Goldberg, 1992).

The underlying theoretical framework of trait extrapolation rests in implicit personality theory (IPT; Jackson, Chan, & Stricker, 1979; Schneider, 1973). IPT was first coined by Bruner and Tagiuri (1954) who proposed that personality traits are relatively fixed over time. This allows individuals to rely on existing knowledge regarding traits, based on experience, to form rapid impressions of others (Dweck, 1999).

A study conducted by Jackson et al. (1979) empirically tested the validity of a rater's IPT existence by correlating an empirically tested list of traits that co-vary with judged co-occurrence of the same traits by participants in the study. The findings revealed that there is a degree of variation in the raters' IPTs, which suggests that individuals differ in recognizing and predicting trait co-variation. Previous research found that trait extrapolation has a moderate effect on predicting rater accuracy (.33; De Kock et al., 2015).

Trait contextualisation. Trait contextualisation refers to the understanding of how certain situations are relevant to specific traits, as previous research shows that certain traits are manifested in specific situations (De Kock et al., 2015; Tett & Guterman, 2000). In other

words, there are certain situations that elicit certain traits to be portrayed. Figure 2 portrays the link between traits and situations within dispositional reasoning. A judge with high trait contextualisation ability comprehends which situations elicit a specific trait.

The theoretical origin of trait contextualisation rests in trait activation theory (Tett & Guterman, 2000). This proposes that individuals differ in their tendencies to express behaviours in certain situations. For example, a good judge will have the ability to understand that extroversion will manifest in a target when they are in a social context, as opposed to when they are by themselves. Situations, therefore, either encourage or discourage trait manifestation (Haaland & Christiansen, 2002; Tett & Burnett, 2003; Robinson, 2009). A good judge will be able to take into account the situations when inferring the target's personality traits. An empirical study showed that trait contextualisation has a moderate effect on predicting rater accuracy (.26; De Kock et al., 2015).

Before we proceed to the empirical findings supporting the positive effect of dispositional reasoning on rater accuracy, we would like to highlight that dispositional reasoning can be interpreted alongside the RAM framework model proposed by Funder (1995; 1999; 2012). As discussed above, dispositional reasoning is the ability of an individual to understand how observable behaviours are manifested from personality traits, situations and behaviours (De Kock et al., 2015). This would require individuals to correctly detect and utilise personality and behavioural cues, which are two stages in judgement accuracy as defined by Funder (1995). Therefore, it is credible to suggest that dispositional reasoning is a product of the judgement process required to make accurate ratings.

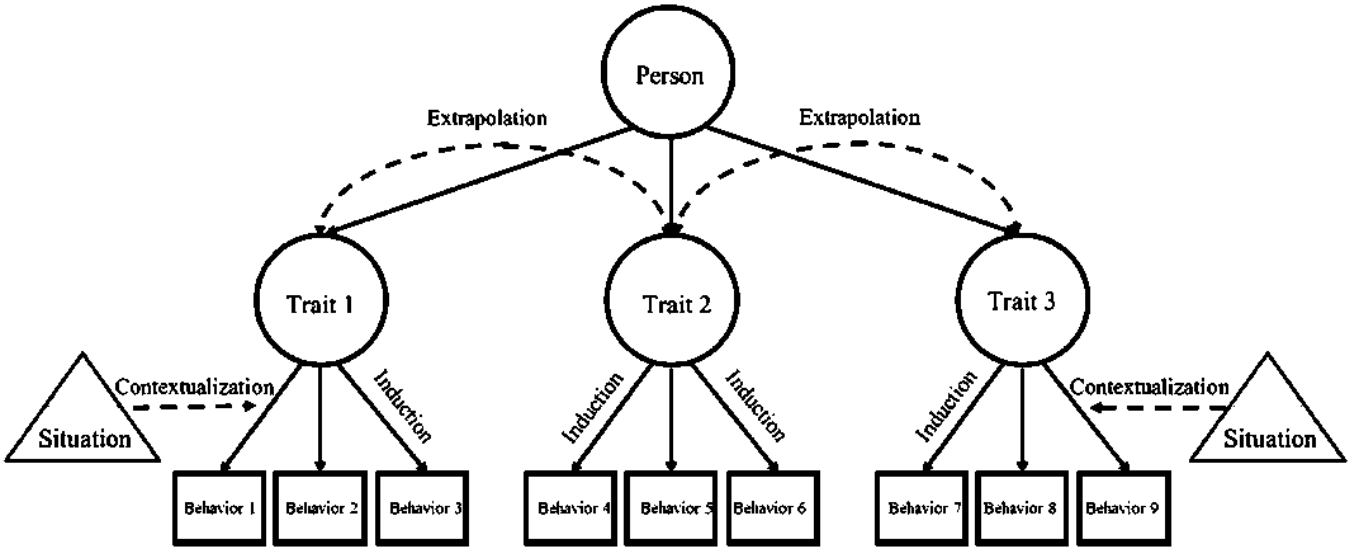

Figure 2. Dispositional reasoning framework by De Kock, F.S., Lievens, F., & Born, M.P., 2015, *Human Performance*, p. 43. Reprinted with permission.

Empirical findings. A seminal study which investigated the effect of dispositional reasoning on rater accuracy was conducted by Christiansen et al. (2005). Findings revealed that dispositional reasoning was the best predicator of rater accuracy (r = .41) amongst other individual characteristics, such as personality and general mental ability. A recent study conducted by Powell and Bourdage (2016) partially replicated these findings and reported that dispositional reasoning predicted participants' ability to accurately judge behaviours presented in the study (r = .22).

Due to the empirical association with rater accuracy, there have been attempts to develop dispositional reasoning and, by extension, rater accuracy through the use of FOR training (see Powell & Bourdage, 2016; Powell & Goffin, 2009). Both studies aimed to investigate whether FOR training would have an effect on dispositional reasoning as well as to explore the causal effect of dispositional reasoning on rater accuracy following the training. Powell and Goffin (2009) revealed that whilst rater accuracy was enhanced in their

study, FOR training had no significant effect on dispositional reasoning. A follow up study conducted by Powell and Bourdage (2016) revealed that those with higher accuracy scores had higher dispositional reasoning scores, however their FOR training efforts had no effect on dispositional reasoning.

A limitation of their findings may rest in their use of a short-term intervention of approximately 30-60 minutes. Longer FOR training sessions may allow for the time needed for dispositional reasoning to be increased, as it is a complex construct (De Kock et al., 2015). Roch et al. (2012) calculated that the average FOR training duration was 100 minutes. The present study intends on extending the length of the FOR training sessions to determine whether a longer FOR training session would result in an effect in dispositional reasoning and consequentially what effect it would have on rater accuracy. Considering the intended extension of training duration and empirical research supporting the predictability of dispositional reasoning on rater accuracy, the following hypotheses are posited:

Hypothesis 2a (H2a): Dispositional reasoning scores will be higher for participants who receive FOR training than participants who receive no training.

Hypothesis 2b (H2b): Dispositional reasoning will be positively related to rater accuracy.

Rater motivation. Rater motivation is a difficult construct to define (Steers & Porter, 1987). For the purpose of this study, rater motivation is defined as the basic goals and objectives that drive the behaviours of raters (Cleveland & Murphy, 1992). Furthermore, it refers to the motivation of the rater to engage in the rating process and to provide accurate ratings (Ispas, 2010).

Several critics of rater accuracy research argued that much of the prior research and its implications have not led to significant improvements in enhancing rater accuracy (Banks

& Murphy, 1985; Ilgen et al., 1993; Roch, 2007; Roch et al., 2012). It is plausible that the reason research and its implications have not led to significant improvements is because research has predominately focused on the rater's ability to increase rater accuracy rather than the role of rater motivation (Banks & Murphy, 1985; Harris, 1994; Yukl, Taber, Longenecker, Gioia, Sims, & Young, 1987). Djurdjevic (2013) argued that it is only recently that researchers have focused on the rater's willingness to provide accurate ratings and that perhaps ratings are strategic decisions made by the rater.

Theoretically, motivation may affect accuracy in multiple ways. One theoretical explanation of rater motivation is offered by the social cognition field (e.g., Fiske & Taylor, 2013). The social cognition field has gained increased attention in explaining the judgement processes involved in rating and how the judgement process effects rating outcomes (Derous, Buijsrogge, Roulin, & Duyck, 2016). In this field, it is credible to suggest that motivation encourages raters either to select conscious judgement processes or automatic and unconscious judgement processes (Fiske & Taylor, 2013).

To elaborate, raters who are high in motivation may report utilising conscious and deliberate judgement processes which are focused on normative rating policies, such as classifying the target's behaviour in a systematic way (De Kock et al., 2018). In contrast, interviewers with lower rater motivation may report utilising unconscious processes that rely on the use of IPTs and other rating heuristics. The use of IPTs and rating heuristics may lead to inaccurate ratings should they be found to be incorrect (De Kock et al., 2018).

A second theoretical explanation of how motivation affects accuracy is the rater motivational model suggested by Harris (1994). Harris postulated that a rater's personal factors (e.g. motivation) and situational factors (e.g. rating context) will most likely influence the rater's attentive and cognitive processes involved in the judgement process, such as the rater's observation of behaviour being rated, information storage, retrieval and memory.

According to Harris (1994), motivated raters apply more attentive and explicit cognitive resources to the rating task. Explicit cognitive processing is characterized by better organization of information and subsequent integration in memory, less use of stereotypes and biases in judgements as well as greater learning (Ispas, 2010), all of which affect the accuracy of ratings (Harris, 1994).

Harris's (1994) argument that motivation increases the rater's attention to behavioural cues and cognitive process can be viewed in light of Funder's (1995; 1999; 2012) RAM framework. It is possible that the increase in attention enhances the rater's ability to detect personality and behavioural cues exhibited by the target, referred to as cue detection. Following this, motivated raters will apply greater cognitive resources and spend more effort in decoding, interpreting and utilising these cues, referred to as cue utilisation. Therefore, by facilitating cue detection and cue utilisation stages, both of which are necessary for accurate ratings (Funder, 1995; 2012), rater motivation may increase the accurate judgement process.

Lastly, it may be credible to suggest that training raters has an effect on their motivation. Prior to receiving training, raters may not find the importance in accurate ratings and would therefore most likely to not be motivated to assign accurate ratings. Roch et al. (2012) suggested that training raters improves 'buy-in' from the raters with regards to the importance of accurate ratings. This 'buy-in' may consequentially increase their motivation to provide accurate ratings.

Empirical findings. Whilst the theoretical underpinnings of rater motivation is promising, there is a dearth of empirical research focusing on the direct relationship between rater motivation and accuracy (De Kock et al., 2018; Spence & Keeping, 2013). Prior research on rater motivation has largely focused on the antecedents of motivation and the factors that influence rater motivation (De Kock et al., 2018). For example, researchers have postulated that motivation may be affected by the belief held by the rater of the anticipated

outcome of the ratings they provide (Vroom, 1964). Other antecedents investigated by researchers include the influence of incentives (Salvemini, Reilly, & Smither, 1993), accountability (Mero & Motowidlo, 2003; Wood & Marshall, 2008) and the use of rater teams (Roch, 2007) on rater motivation.

Contemporary empirical research has not expanded on the direct relationship between motivation and on rater accuracy (De Kock et al., 2018), nor on the influence of motivation on FOR training. This further encourages the need to investigate whether rater motivation influences the effectiveness of FOR training in enhancing accuracy. Based on this suggestion for future research and the promising theoretical framework surrounding rater motivation, the present paper will investigate whether training interviewers will increase their levels of motivation and investigate the relationship between rater motivation and rater accuracy. Based on the theoretical arguments on rater motivation, it is hypothesised that:

Hypothesis 3a (H3a): Participants who receive FOR training will have higher self-reported motivation than participants who receive no training.

Hypothesis 3b (H3b): Rater motivation will be positively related to rater accuracy.

Rater self-efficacy. Self-efficacy is defined by Bandura (1986) as an individual's sense of personal mastery and a judgement of their ability to perform an action to achieve a particular outcome. Rater self-efficacy refers to an individual's belief and confidence in their ability to execute the behavioural demands of the rater role (Bernardin & Buckley, 1981). This belief depends on their evaluation of whether they are able to successfully carry out the range of sub-tasks involved in the rating task in order to provide ratings that are perceived as fair and accurate (Bernardin & Buckley, 1981; Wood & Marshall, 2008).

Rater self-efficacy is favourable in the rating context, however some raters may experience inefficacy. In other words, some raters lack the belief that they are able to

effectively execute the role of a rater (Wood & Marshall, 2008). This inefficacy arises from various sources. For example, some raters may doubt their ability to rate accurately or feel that they do not understand the target's job and behaviour, and thus believe they cannot truly execute the task of rating that said job or behaviour. Other sources may include raters feeling uncomfortable during the rating process or doubting their interpersonal skills and knowledge necessary when conducting ratings in various contexts. Raters who lack efficacy are more likely to reduce their efforts, become less systematic when processing information and are more likely to provide lenient ratings to become more comfortable in the rating process (Bandura, 1997; Benedict & Levine, 1988; Wood & Bandura, 1989), all of which will affect the accuracy of their ratings.

It has been argued that the rating experience and training raters alone will not equate to more effective and accurate ratings (Wood & Marshall, 2008). Central to accurate ratings is a sense of personal mastery and the rater's belief in their ability (Sedikides & Skowronski, 1991). Therefore, rater training and its effect is largely dependent on the rater's belief that they are able to handle difficulties and problems as they arise (Lievens, 2001). Rater training needs to develop the rater's belief that they are able to transfer the skills gained in the training to the rating context, otherwise it is argued that the training will be ineffective in enhancing rater accuracy (Lievens, 2001). Consequently, should they hold high levels of self-efficacy, raters will apply greater effort to override difficulties and problems in the rating context by using the skills and knowledge gained in training (Wood & Marshall, 2008).

This effort in overriding difficulties by using the knowledge and skills gained from the training can be further interpreted using Funder's (1995; 1999; 2012) RAM framework. Should raters who have a high sense of self-efficacy apply greater effort in overriding difficulties by using their skills and knowledge of the rating task, then it is plausible that they will be able to apply greater cue detection and cue utilisation. Cue detection requires the rater

to correctly detect behavioural and personality cues during the rating process. Cue utilisation requires the rater to correctly use the cues to make an evaluation (Funder, 1995; 1999; 2012). Therefore, should a rater have difficulty in utilising a cue that has been made available to them, raters with higher self-efficacy would apply greater effort in overcoming these difficulties and proceed to use their skills and knowledge to correctly detect and utilise the cues.

Empirical findings. A key study supporting the belief that rater self-efficacy has an influence on rater accuracy was conducted by Wood and Marshall (2008). The researchers measured participants' (N = 194) ratings of a video portraying a nurse's behaviour which they then compared to expert ratings. They also measured the participant's self-reported self-efficacy scores through a self-developed measure (PASE; Wood & Marshall, 2008). They reported that rater self-efficacy was positively related to rating accuracy (r = .39).

In addition, a study conducted to investigate the influence of self-efficacy in FOR rater training found FOR training increased self-efficacy amongst participants who received the training (Dierdorff et al., 2010). It is noted that whilst Dierdorff et al., (2010) conducted a similar study to the present study—in that they explored the effect of self-efficacy in the FOR context—they focused on learning self-efficacy. Learning self-efficacy differs from rater self-efficacy, in the sense that the former focuses on the belief and confidence of the raters in their ability to learn from the training programme (Dierdorff et al., 2010). The present study however focuses on latter, namely the rater's belief in their ability to rate accurately following the training.

Whilst these findings seem promising, rater self-efficacy remains relatively unexplored (De Kock et al., 2018). The present paper will attempt to expand further on earlier work by examining the influence of rater self-efficacy on rater accuracy in a FOR training context. If a rater has a high resilience and confidence in their knowledge and skills gained in

the FOR training, they are most likely to make difficult decisions following the rating, which will assist in the accuracy of their ratings (See Figure 3; Wood & Marshall, 2008). Therefore, it is hypothesised that:

Hypothesis 4a (H4a): Participants who receive FOR training will have a higher sense of rating self-efficacy than participants who receive no training.

Hypothesis 4b (H4b): Rater self-efficacy will be positively related to rater accuracy.

To summarise, the independent variable in the present study is the FOR training approach and the dependent variable is rater accuracy. It is proposed that FOR training will positively affect rater accuracy (H1; See Figure 3). In addition, it is proposed that FOR training will positively affect dispositional reasoning (H2a), rater motivation (H3a) and rater self-efficacy (H4a) and that these three variables will positively influence rater accuracy (2b, 3b, 4b). To test the above mentioned hypotheses, the researcher followed an experimental approach in line with existing FOR training and rater accuracy literature.

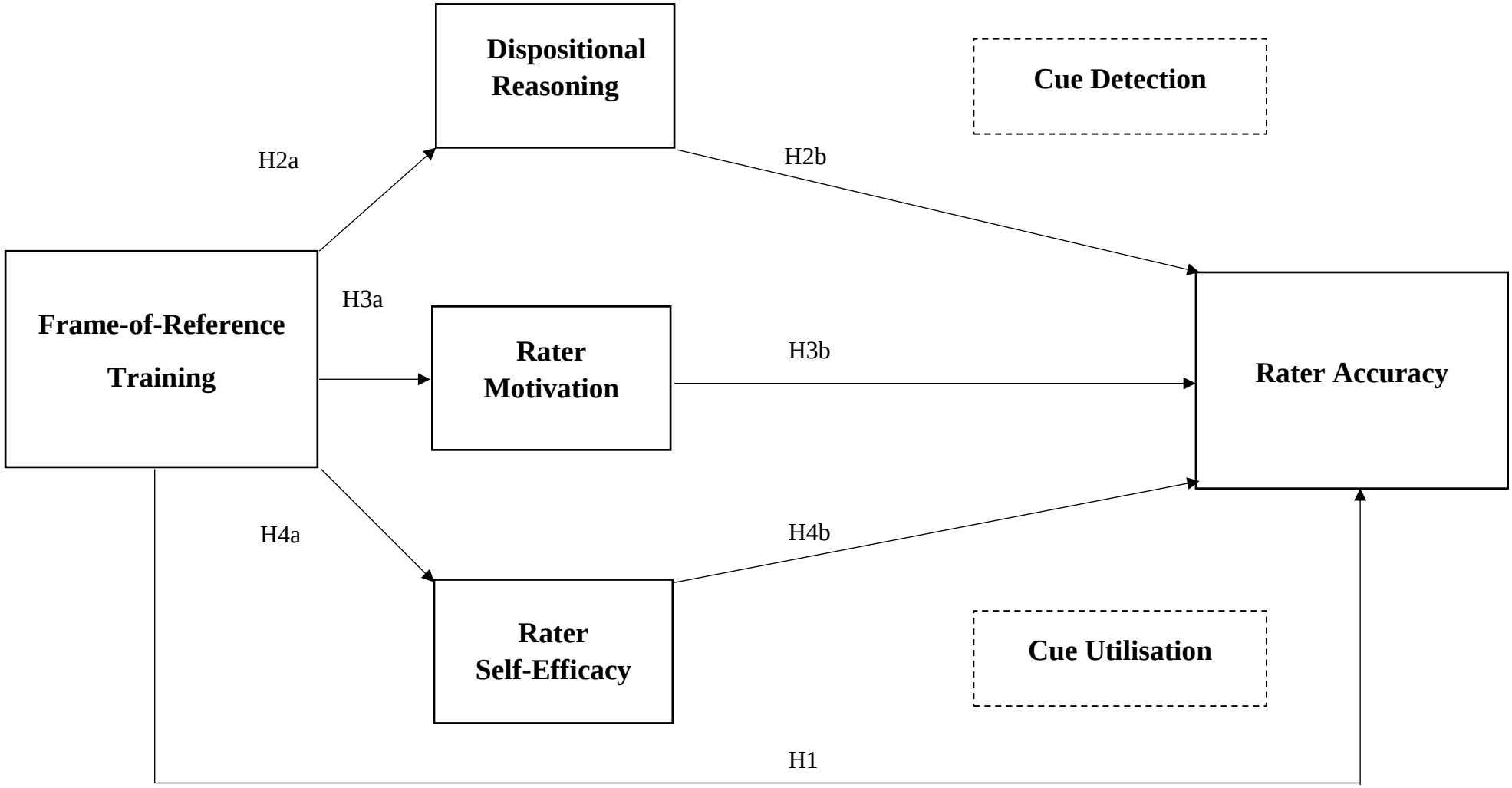

Figure 3. The integrated conceptual model representing the theoretical framework for the present study.

Method

Research Design

The study used a post-test only experimental research design with random assignment of participants (see Table 1; Campbell & Stanley, 2015). Participants were randomly crossed with either the FOR training condition or the no-training condition, which acts as our control group. By using this research design, the researcher would be able to deduce casual inferences on whether the independent variable, namely FOR training, had any effect on the dependent variables in the study, namely motivation, self-efficacy, dispositional reasoning and rater accuracy. The decision was made not to include a pre-test in the design in order to control for test maturation and potential learning effects (Burns & Burns, 2008; Cozby, 2007). The design is well suited to address the research question, namely to test and assess the influence of our study's variables on the effectiveness of the FOR training approach in enhancing rating accuracy (Campbell & Stanley, 2015; Millsap & Maydeu-Olivares, 2009). Due to time constraints, a cross-sectional design was used.

Table 1

Research Design

Treatment group	Intervention	Observation
R_1	X	O
R_2		O

Note. R_1 = experimental training group; R_2 = no-training group; X = frame-of-reference training intervention; O = post-test.

Participants

Non-probability convenience sampling was used to source participants. In line with previous FOR studies (Lievens, 2001; Powell & Bourdage, 2016; Powell & Goffin, 2009), students were used in the sample. In addition, students have been argued to be more readily able to adopt the well-established schemas imposed by training (Schleicher & Day, 1998). To increase external validity, the researcher recruited all participants from a final year Organisational Psychology course within the School of Management Studies at a university in South Africa. By sourcing participants from the Organisational Psychology course, participants would likely have had prior knowledge of interviews and the selection process gained through the courses required for their qualification. To increase participation, data were collected on three separate occasions to accommodate students who differed in availability. Three lucky draw research participation incentives of R350 were offered in order to further increase participation (see Harris, 1994). These incentives were paid out in cash to the winners a month after data collection.

A total of 32 participants were recruited and were randomly assigned to either the training condition group (n = 16) or the control group (n = 16) by issuing them with a random number upon arrival that was pre-allocated to represent either the FOR training condition or no-training condition (Burns & Burns, 2008). Participants were young adults between 20 to 28 years old (FOR training: M = 21.81, SD = 1.22; No training: M = 21.56, SD = 1.67). Table 2 below contains language, gender and race demographic statistics for both conditions. It is noted that the sample consisted largely of White English speaking female students. This may have affected the results which is problematic for the generalisability to the larger population of raters (Rosenthal & Rosnow, 2008).

Table 2

Demographic Sample Characteristics (N = 32)

		Training condition			
		FOR (n = 16)		NT (n = 16)	
Variables		*f*	%	*f*	%
Gender	Female	14	87.5	14	87.5
	Male	2	12.5	2	12.5
Race	White	8	50	4	25
	Black	4	25	5	31.3
	Indian	1	6.3	2	12.5
	Coloured	2	12.5	5	31.3
	Other	1	6.3	0	0
Language	English	13	81.3	10	62.5
	Xhosa	0	0	4	25
	Zulu	1	6.3	0	0
	Sotho	1	6.3	0	0
	Korean	1	6.3	0	0
	Siswati	0	0	1	6.3
	Shona	0	0	1	6.3
Year of study	3rd year	11	68.8	10	62.5
	4th year	5	31.3	6	37.5

Note. FOR = frame-of-reference training; NT = no-training; *f* = frequency.

Materials

Interview construction. Due to time constraints, videotaped segments of previously recorded interview performance were used as stimuli in the present study, as developed by De Kock et al. (2015). In their study, the researchers video-recorded semi-structured interviews of graduate students recruited to take part in an interview that would help them prepare for the job application process.

The interview format was a competency-based, situational interview (Latham, Saari, Pursell, & Campion, 1980) and an expert interviewer conducted each of the interviews in the same manner and structure. In each interview, four interview questions were posed in the same order to target the people management dimension. This dimension was selected given its widespread use in interviews (Huffcutt, Conway, Roth, & Stone, 2001) and applicability to the fictional position of a Trainee Manager used as the job context for the interviews.

Each video-recorded interview contained only the audio and visual of the graduate student being interviewed and not the expert interviewer. Due to the fact that the interviewer's audio or visual were not present in the recorded interviews, the questions posed by the interviewer were presented in text in the recordings before the graduate responded to the questions. Four video-taped recordings were selected and were shortened to a viewing time of approximately three minutes each.

Video interviewee true scores. In line with earlier recommendations for true score estimation (Sulsky & Balzer, 1988), De Kock et al. (2015) developed the true scores through combined ratings from multiple subject matter experts (SMEs), comprising of qualified industrial psychologists and lecturers in I-O psychology. The authors asked a panel of seven SMEs to rate the videotaped applicants on the people management dimension. As per Borman's (1977) procedure for true score estimation, the authors allowed the panel to view the recorded applicants as often as they wished before completing the structured rating sheet

(De Kock et al., 2015). The panel was reported to be balanced in terms of gender and ethnicity in order to minimize possible demographic effects. To obtain overall true score estimates for each interviewee on the performance dimension, De Kock et al. (2105) averaged the ratings made by the respective SMEs.

Development of training protocol and training material. The researcher adapted a well-cited FOR training and no-training protocol developed by Lievens (2001). Firstly, due to time constraints, the researcher adapted the protocol by shortening information in the workshop session regarding the interview process. Secondly, the training protocol was adapted by including only people management information used in Lieven's (2001) training.

The researcher supplemented the training content pertaining to the people management dimension with relevant behaviourally anchor rating scales (BARS) developed by De Kock et al. (2015), as well as behaviours relating to interpersonal styles from the Riverside Behavioural Q-Sort (RBQ; Funder, Furr & Colvin, 2000). For example, one of the BARS developed by De Kock et al. (2015) indicated an effective behaviour of people management was "to set the example, whilst providing incentives" in order to motivate employees to work harder. Behaviours that correlate with unsuccessful interpersonal style, according to the RBQ, was "exhibits awkward interpersonal style". These were added to the training content pertaining to people management.

With regards to training material, the FOR training condition was given a paper-and-pencil exercise as a training exercise. The task consisted of eight items which contained a brief example of people management behaviour. Participants were required to assign a category of effective, average and ineffective indications of people management to the behaviour example. Similar to the training content, the training exercise content was based on the BARS developed by De Kock et al. (2015) as well as the RBQ (Funder, Furr & Colvin, 2000).

Finally, Microsoft PowerPoint was used to create sets of lecture slides that would assist trainers in providing the training. The lecture slides were designed to aid in structuring the training. To assess instruction clarity, item difficulty and time adequacy of both training conditions, a pilot test was conducted (Burns & Burns, 2008).

Data Collection and Procedure

Data was collected on three separate occasions over a one-week period during September 2017. The relevant university's Ethics in Research Committee granted permission to undertake the research (see Appendix A). The executive director of student affairs at the university granted access to the student sample. Participants had the right to withdraw from the research at any time and were required to indicate consent prior to participation in the research (see Appendix B).

To control for experimenter demand effect, which is the possibility of behavioural changes in participants due to cues made by the researcher of what constitutes acceptable behaviour (Finkelstein, 1976; Rosenthal, Kohn, Greenfield, & Carota, 1966; Valentine, 1992), two Master students conducted the training. The trainers were not involved in the present study further than delivering the training. The students were trained in both the FOR condition as well as the no-training condition and they were randomly assigned to a training condition in the same manner as participants (Burns & Burns, 2008). Trainers were not informed about which conditions they had been assigned to.

Rater training. Participants gathered at an assigned venue on campus. Consent was obtained from each participant before the experiment began. Participants in both the experimental group and in the control group participated in an introductory workshop. In line with the earlier FOR studies (Lievens, 2001; Powell & Goffin, 2009), the workshop consisted of three main components. Firstly, a brief lecture was given about the basics of interviews including the purpose, components and current usage of interviews. Secondly, the

participants were given information regarding the fictitious working context (e.g. tasks, duties and required qualifications) of the Trainee Manager position and the organisational context. Finally, the dimension of people management was presented to participants. Following the workshop, participants split into the two training conditions and were instructed to relocate to the assigned venues, depending on their numbers randomly assigned to them upon arrival.

FOR training session. As conducted in previous FOR studies (Lievens, 2001; Powell & Bourdage, 2016; Powell & Goffin, 2009; Woehr, 1994), the trainers presented a definition of the people management dimension and then provided the participants with examples of effective, average and ineffective behaviours that relate to the dimension. Participants were instructed that they could use this information to scan the behaviours presented in the videotaped interviews.

Participants were then given a written exercise that listed eight incidents portraying a specific behaviour related to people management. Participants needed to assign each incident to a respective performance category (effective, average and ineffective). After completing the written exercise, participants were instructed to divide into groups and discuss their reasons for their assigned category for each of the eight incidents. Following this discussion, the trainers then discussed the participants' answers and provided feedback as to the correct category assignments for each of the eight incidents.

Finally, participants practiced their ratings by viewing and evaluating a videotape that portrays a candidate being interviewed for the Trainee Manager position. Thereafter, the trainer conducted a discussion session on how the participants decided to assign a rating to the candidate and clarified any differences in ratings amongst the group. The trainer then provided participants with the feedback regarding their ratings and the true scores of each candidate.

No-training session. Following previous FOR studies (Lievens, 2001; Powell & Bourdage, 2016), the control condition involved a practice rating session. No specific training concepts related to people management were relayed to the control group, however, nor did they participate in the written exercise. Participants were simply instructed to watch the practice video and thereafter provide ratings. Following the practice video, the participants were then instructed to divide into groups and discuss their assigned ratings. No feedback was given regarding the true scores of the practice video.

Although this condition served as our control condition, the participants were made to believe they were being trained, through the practice rating and discussion session, which overcomes the limitations of pure no-training control conditions (Cook, Campbell, & Day 1979). Cook et al. (1979) suggested that individuals provided with no training would presume they are in the control group and would have low motivation to make accurate ratings, which would have affected our results.

Final interview rating session. Following the training sessions, participants in both the training and no-training condition groups were instructed to observe recorded performances of three candidates that were being considered for the position of Trainee Manager. For recording behaviours of each of the candidates, participants were provided with observation forms to take notes. Following the observation of each candidate, participants were instructed to rate the candidates on each of the four questions using a 7-point scale ranging from 7 = “excellent”, 4 = “moderate” to 1 = “poor”.

Debriefing. Upon completion of the experiment, participants were fully debriefed in person on the nature of the study by the researcher during the participant’s lecture time. Participants were made aware of the two training conditions, the study variables, how their results may be used by the researchers and were provided with the opportunity to ask

questions regarding the study. Participants were then asked to provide their email addresses if they would like to receive feedback on their test scores.

Measures

Rater accuracy. Consistent with previous research studies (e.g., Christiansen et al., 2005; De Kock et al., 2015; Powell & Bourdage, 2016; Powell & Goffin, 2009), the accuracy score was computed for each participant by calculating within-person profile correlations. Borman's Differential Accuracy (BDA; 1977) was used to explore the correlation between the participant's overall ratings and the corresponding true score. Higher scores on the BDA will reflect higher accuracy (Borman, 1977; Sulsky & Balzer, 1988). Correlations were transformed using an r-to-Fisher's-z transformation.

Dispositional reasoning. An adapted and shortened form of the Revised Interpersonal Judgement Inventory (RIJI; De Kock et al., 2015) measured dispositional reasoning. A shortened form of the RIJI was necessary due to the need to keep the overall length of the measures as short as possible to avoid drop out and respondent fatigue (Burns & Burns, 2008; Cozby, 2007). The test consists of three subscale measures, namely: trait induction, trait extrapolation, and trait contextualisation measures. The RIJI showed evidence of construct validity and differential prediction of the components on the accuracy criterion, specifically: trait extrapolation (.33), trait contextualisation (.26) and trait induction (.14; De Kock et al., 2015; De Kock et al., 2018). Moreover, they demonstrated discriminant validity with personality and incremental validity over cognitive ability in predicting interview rater accuracy (De Kock, Lievens, & Born, 2017). All three subscales showed acceptable confirmatory factor analysis (CFA) derived construct reliabilities (induction = .77; extrapolation = .81; and contextualisation = .76; De Kock et al., 2015).

Items used in the present study were selected by assessing confirmatory factor analysis factor loadings from De Kock et al.'s (2015) study, using the full RIJI version on a

combined sample of psychology students and managers (N = 321). Items were selected based on the highest factor loadings, leaving 18 items in the final measure.

Trait induction. This subscale measures behaviour-trait inferences (De Kock et al., 2015). The subscale describes each of the Big-Five personality traits and then requires participants to identify which traits (e.g. conscientiousness) are best suited to a list of adjectives (e.g. thorough) from Goldberg's (1992) factor markers. Based on factor loadings (De Kock et al., 2015), ten items were selected.

Trait extrapolation. This subscale measures the understanding of trait co-occurrence (De Kock et al., 2015). In each item, a fictitious "paper person" is presented. Thereafter, participants are required to select one of four descriptions that is most likely to also be true of said fictitious person. Based on factor loadings (De Kock et al., 2015), four items were selected.

Trait contextualisation. This last subscale measures the understanding of trait-situation relevance, in other words how situations are related to trait occurrence (De Kock et al., 2015). This subscale is divided into two subsets. The first subset presents a trait description by listing examples of behaviours related to high and low scores of the trait, and then requires the participants to choose which situations would most likely provoke the relevant behaviour described. The second subset describes a situation and then requires participants to identify the trait most likely to be observed in that specific situation. Based on factor loadings (De Kock et al., 2015), four items were selected.

Rater motivation. An adapted form of a previously developed rater motivation measure (Hedge & Teachout, 2000) was used to measure participants' motivation in the study. This measure was used because it was short in nature, it has been used in previous similar studies (e.g., Ispas, 2010; Roch, 2007; Roch, McNall, & Caputo, 2011), and was found to be high in reliability (α = .91; Hedge & Teachout, 2000).

The measure was adapted by alternating one of the items to become a reverse item and combining two similarly worded items into one item (e.g. “important to make accurate ratings” and “in general, accurate ratings important”). There were seven items in total and participants rated each of the items on a 5 point Likert-type response scale ranging from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”). After participants completed the measure and the reverse item was coded, motivation scores were computed as the participant’s mean item response by adding all the responses and dividing it by the number of items in the measure.

Rater self-efficacy. An adapted form of the well-cited Performance Appraiser Self-Efficacy (PASE; Wood & Marshall, 2008) scale was used to measure participants’ rating self-efficacy. The PASE scale was used as it is short in nature, it has been used in a recent study (see Moser, Kemter, Wachsmann, Kover & Soucek, 2016) and was found to have high reliability ($\alpha = .85$; Wood & Marshall, 2008). By assessing factor loadings in Wood and Marshall’s study (2008), five items were selected.

The measure was adapted by rephrasing items to be more suitable for the interview context as opposed to the performance appraisal context, which was PASE’s original context (Wood & Marshall, 2008). For example, the original item “explain to persons of higher authority the reasons for assigned ratings” was adapted to “explain to trainer the reasons for assigned ratings”. The adaptation was done per Bandura’s (2006) guidelines in constructing self-efficacy scales, namely to phrase items to say “can do” rather than “will do” and to pre-test the adapted items. Participants responded to items (including a reverse item) on a 5 point Likert-type scale ranging from 1 (“Strongly Disagree”) to 5 (“Strongly Agree”). After participants completed the measure and the reverse item was coded, self-efficacy scores were computed as the participant’s mean item response by adding all the responses and dividing it by the number of items in the measure.

Personality. Personality was measured in order to determine whether the different samples in each training condition differed in personality. This may have affected the accuracy scores. The Big Five Inventory- 2-Short Form (BFI-2- S; Soto & John, 2017) scale was used to measure participants' personalities on Extraversion, Agreeableness, Contentiousness, Emotional Stability and Open-Mindedness. The BFI-2- S is a shortened version of the 60 items Big Five Inventory-2 measure. The BFI-2- S contains 30 items that ask the participant to view statements that reflect personality characteristics, and to rate how applicable the statements are to the participant. The items are measured on a 5 point Likert-type scale ranging from 1 ("Strongly Disagree") to 5 ("Strongly Agree"). The BFI-2-S has an alpha reliability coefficient of .78 and retains 80 percent of the full measure's reliability and validity properties (Soto & John, 2017). Thus, the decision was made to use the shortened version to minimise assessment time and fatigue (Burns & Burns, 2008).

Intelligence. Similar to the need to measure personality, the participant's intelligence was also measured in each condition. The International Cognitive Ability Resource sample test (ICAR16; Condon & Revelle, 2014) was used to measure participants' general intelligence. The ICAR sample test contains 16 items and is a shortened version of the ICAR 60 item cognitive ability measure (Condon & Revelle, 2014). The 16 items range across four item types: matrix reasoning, letter and numerical reasoning, 3D rotation reasoning and verbal reasoning tests. The ICAR sample test is high in reliability ($\alpha = 0.81$) as well as in validity (Condon & Revelle, 2014), and thus the decision was made to use the shortened version due to time constraints.

Biographical. Participants completed a biographical questionnaire that recorded gender, age, race, first language and year of study for statistical purposes.

Manipulation checks. A two item measure was used to assess the consistency of training delivery by each trainer, which could have affected the results of each training

condition (Burns & Burns, 2008): “I felt that the trainer was professional” and “I felt that the trainer was enthusiastic”. An additional two item measure was used to assess the training manipulation between the two training conditions: “I had a good idea of what the performance dimension was” and “I know what to look for in the interviews following the training”.

All measures used in the study, with the exception of the intelligence measure, were paper-and-pen based and were administered directly following the training intervention. This decision was made for practical purposes based on the limited venue and computer availability needed for all the measures to be completed online. The intelligence measure was administered via Qualtrics (Version 37,892), a week following the training intervention and drop out was avoided by instructing that only participants who completed the full intelligence measure would qualify to receive the research participation incentive.

Statistical Analysis

The researcher scored and coded the paper-and-pencil scores into Microsoft Excel. The data was checked on two separate occasions to ensure no errors were made by the researcher when transferring the paper-based scores into Excel. Thereafter, the researcher imported the Excel spreadsheet to IBM Software Package for Social Sciences (SPSS; Version 22). The ICAR data were scored and coded directly into SPSS. Descriptive statistics were used to describe the sample. Due to the small sample size, non-parametric statistics were used to test the hypotheses, specifically the Mann-Whitney U tests to compare the two training conditions on the study’s variables and the Kendall’s Tau correlation tests to investigate the relationship between our study variables (Field, 2013). Ideally, a MANOVA test would have been conducted to assess the mediation effect of the study’s variables on the dependent variable, however due to the small sample size, it was not feasible to conduct parametric statistics and test for mediation effects (Field, 2013).

Results

Measurement Properties

For the purpose of this study, internal consistency and dimensionality analysis for the measurements used was precluded because of the small sample size (De Bruin, 2004; Nunnally & Bernstein, 1978). Furthermore, exploratory and confirmatory factor analysis could not be undertaken. As previously discussed, prior empirical studies have supported the measurement properties of all measures used in this study.

Preliminary Analyses

Invalid cases. Prior to further analysis of the obtained data from 32 participants, the data set was cleaned and explored for invalid or missing cases (Field, 2013). When exploring for missing cases in each training condition it was found that there were none, as all participants completed each measure and no participants dropped out. It was decided therefore to include all participants, thus leaving the study with 32 participants.

Normality. The normality assumption of the variables in the present study was explored using the Shapiro-Wilk test as the sample sizes were small for each training condition (Field, 2013). A significance value which was equal to or greater than .05 was used to indicate normality for each training condition in relation to each of the relevant variables. The FOR training condition had normality for each variable as the *p*-values were above .05 (See Table 3). The no-training condition revealed *p*-values greater than .05 for both accuracy and dispositional reasoning but not for motivation and self-efficacy. This suggests that the data were not normal for rater motivation and rater self-efficacy in the no-training condition. Based on the results of this assumption, as well the small sample size, it was decided to conduct non-parametric methods for the analyses (Field, 2013).

Homogeneity of variance. Equivalence of variance across conditions is a necessary assumption for non-parametric statistics (Field, 2013). A Levene's test was conducted to test

for the assumption of homogeneity of variance (Field, 2013). Equal variance was found between the FOR training condition and the no-training condition (F (1, 30) = 1.04, p = .32, n.s.). It was thus appropriate to conduct an independent sample t-test to test the hypotheses.

Table 3

Shapiro-Wilk's Test of Normality of Data: Frame-of-Reference and No-Training Conditions

Condition	Variable	*W*	*df*	*p*
FOR	Accuracy	.98	16	.93
	Motivation	.90	16	.08
	Self-Efficacy	.94	16	.35
	Dispositional Reasoning	.94	16	.46
NT	Accuracy	.96	16	.61
	Motivation	.89*	16	.05
	Self-Efficacy	.87*	16	.03
	Dispositional Reasoning	.96	16	.57

Note. FOR = frame-of-reference training; NT = no-training

$p < .05^{*}$ (two tailed)

Descriptive Statistics

The means, medians and standard deviations of the study variables are presented below in Table 4. Results are presented separately by condition, namely FOR training and no-training. On average, participants in the FOR training condition showed considerably greater accuracy scores (M = .68, SD = .32) than the no-training condition (M = .15, SD = .39). In addition, participants in the FOR training condition scored higher on dispositional reasoning (M = 15.38, SD = 1.46), rater motivation (M = 4.60, SD = .23) and rater self-efficacy (M = 4.30, SD = .36) compared to participants in the no-training condition. The standard deviation was most noticeable between the FOR training and the no-training condition on dispositional reasoning scores, suggesting that

scores on this variable were more spread out around the mean. The standard deviation for accuracy, motivation and self-efficacy indicated a narrow distribution around the mean as the standard deviations were small values equal to or below .54.

The data set was analysed for skewness, which investigates the symmetry of the distribution of scores, as well as kurtosis, which investigates the degree to which scores cluster at the ends of distribution (Burns & Burns, 2008; Field, 2013). Distribution is considered to be normal when the skewness and kurtosis values are zero, or relatively close to zero (Burns & Burns, 2008). In the FOR training condition, all study variables indicated a moderately negative distribution with skewness ranging from -.53 to -.14. This was found to be similar in the no-training condition, with skewness negatively ranging from -.86 to -.45, with the exception of self-efficacy having a positive skewness of .09. Further examination of the kurtosis value for the FOR training condition, ranging from -.83 to 8.2, and the no-training condition, ranging from -1.59 to -.28, show that these are all below 3, thereby indicating a platykurtic distribution (Field, 2013).

Table 4

Descriptive Statistics

Condition	Variable	*n*	*Mdn*	*M*	*SD*	Min	Max	95% CI
FOR	Accuracy	16	.70	.68	.32	-.06	1.27	[.51, .85]
	Motivation[a]	16	4.60	4.60	.23	4.14	4.86	[4.48, 4.73]
	Self-Efficacy[b]	16	4.20	4.30	.36	3.6	4.8	[4.11, 4.49]
	Dispositional Reasoning[c]	16	15	15.38	1.46	13	18	[14.60, 16.15]
NT	Accuracy	16	.18	.15	.39	-.69	.69	[-.06, .35]
	Motivation	16	4.57	4.43	.49	3.42	5	[4.17, 4.69]
	Self-Efficacy	16	4.00	4.03	.54	3.4	4.8	[3.74, 4.31]
	Dispositional Reasoning	16	13.50	13.13	2.66	8	17	[17.71, 14.54]

Note. FOR = frame-of-reference training condition; NT = no-training; CI = confidence interval.

[a] Scores on the motivation scale have a possible range of 1 to 5.

[b] Scores on the self-efficacy scale have a possible range of 1 to 5.

[c] Scores on the dispositional reasoning measure have a possible range of 0 to 18.

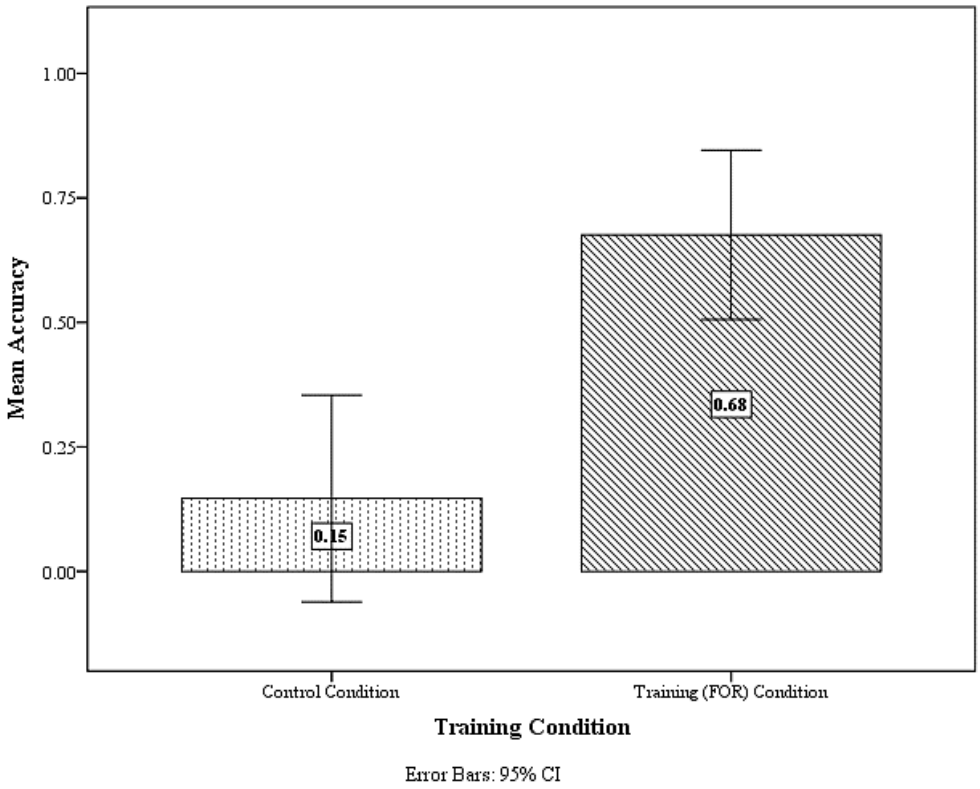

Figure 4. Mean rater accuracy scores across FOR training and no-training conditions.

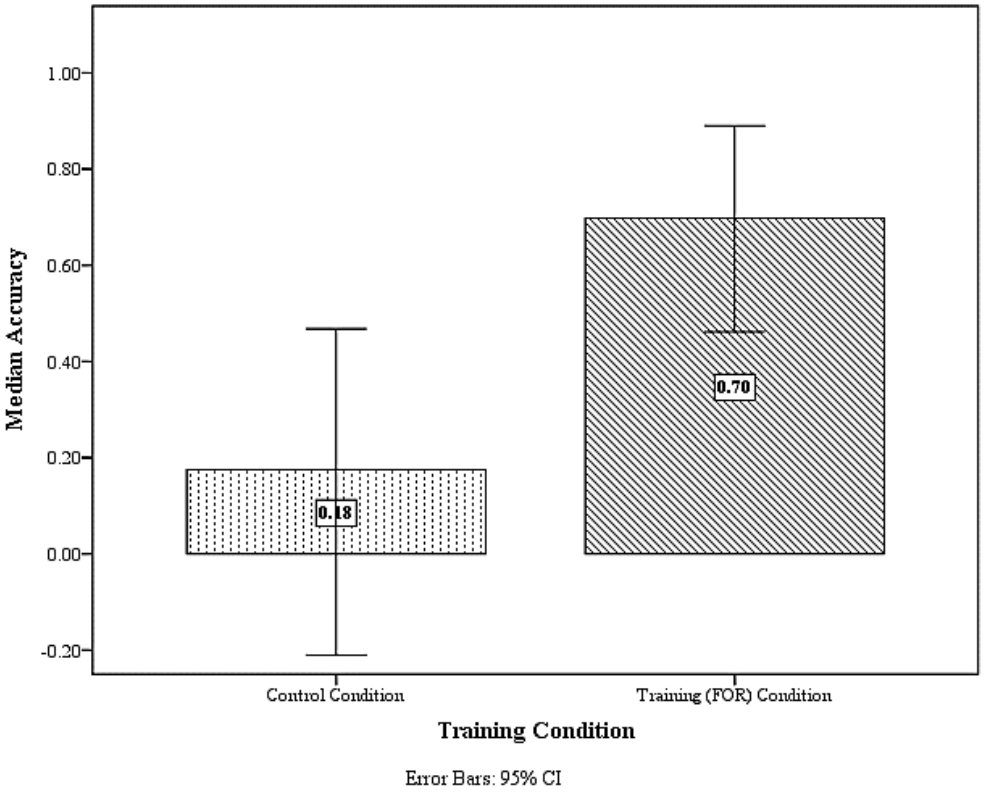

Figure 5. Median rater accuracy scores across FOR training and no-training conditions.

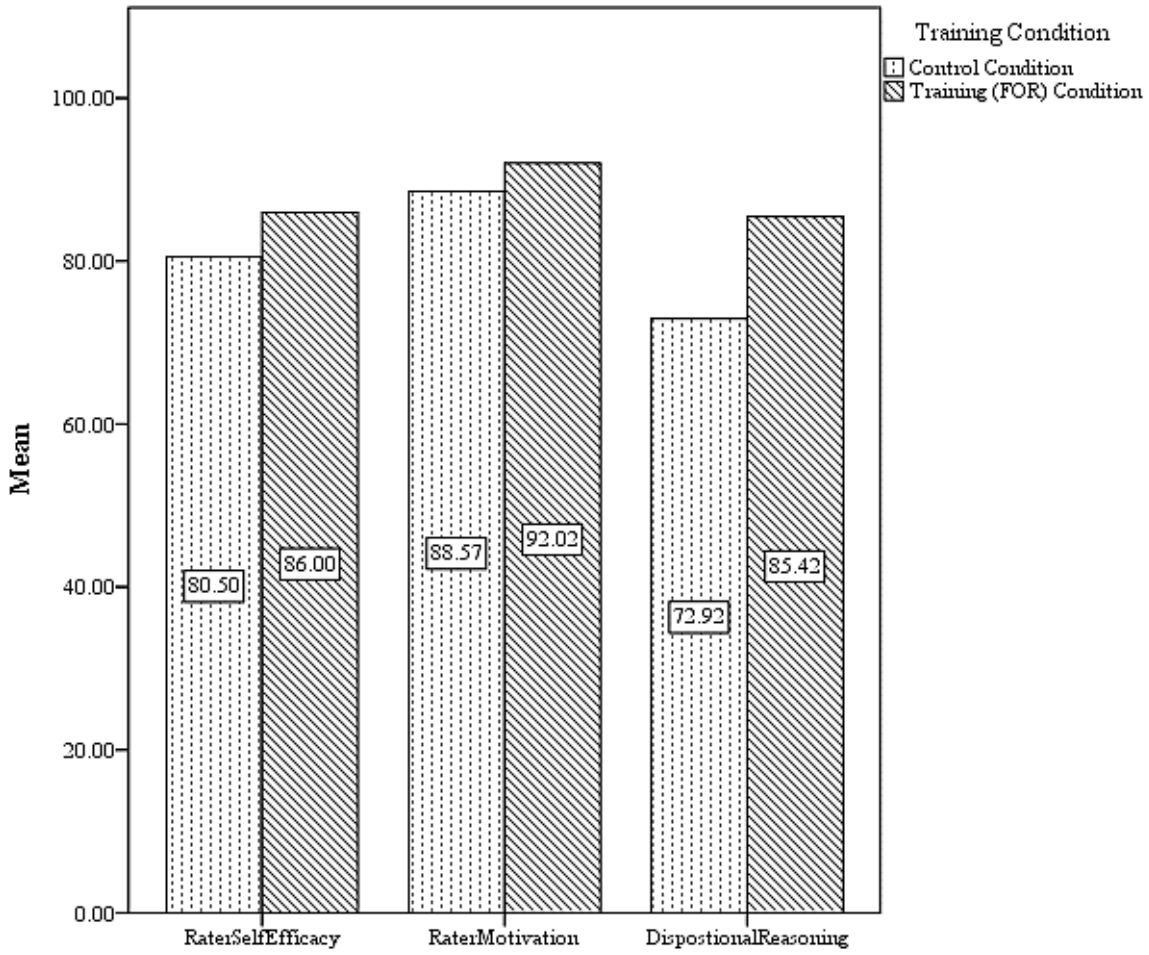

Figure 6. Mean percentages of self-efficacy, motivation and dispositional reasoning across the FOR training and no-training conditions.

Testing of Hypotheses

Hypothesis 1. Hypothesis 1 stated that participants in the FOR training conditions would have higher rater accuracy scores than the no-training condition. A Mann-Whitney U test attempted to determine the statistical significance of the difference between the medians of the two groups (Field, 2013). A significant difference in rater accuracy was found between the FOR conditions (*Mdn* = .68), and the no-training condition (*Mdn* = .18), $U = 223$, $z = 3.58$, $p = .00$. The null hypothesis is, therefore, rejected. Results indicated a large effect size $d = .68$ (Cohen, 1988).

Hypothesis 2a. Hypothesis 2a stated that participants in the FOR training condition would have higher dispositional reasoning scores than the no-training condition. A Mann-Whitney U test revealed a significant difference in rater accuracy was found between the FOR condition (*Mdn* = 15), and the no-training condition (*Mdn* = .13.50), *U* = 193.5, *z* = 2.49, *p* = .01. The null hypothesis is, therefore, rejected. Results indicated a medium effect size *d* = .44 (Cohen, 1988).

Hypothesis 2b. Hypothesis 2b stated that participants who scored higher dispositional reasoning scores would also score higher accuracy scores. A Kendall's Tau test was used to determine the relationship between dispositional reasoning and rater accuracy. This is a non-parametric correlation method appropriate for smaller sample sizes and when the data has many values with the same score (Field, 2013). Results indicated that whilst there was a positive relationship between dispositional reasoning and accuracy scores, the relationship was insignificant (r_τ = .17, *p* = .19). This lack of significance may be due to the small sample size; however, the null hypothesis is retained.

Hypothesis 3a. Hypothesis 3a stated that participants in the FOR training conditions would have higher self-reported motivation scores than the no-training condition. A Mann-Whitney U test found no significant difference in motivation between the FOR condition (*Mdn* = 4.60) and the no-training condition (*Mdn* = 4.57), *U* = 144, *z* = .61, *p* = .56. Results indicated a small effect size *d* = .11 (Cohen, 1988). The small sample size could be the reason for the non-significant result. The null hypothesis is retained.

Hypothesis 3b. Hypothesis 3b stated that rater motivation would be positively related to rater accuracy. A Kendall's Tau test was used to determine the relationship between motivation and accuracy, which indicated that whilst there was a positive relationship between motivation and accuracy scores, the relationship was insignificant (r_τ = .17, *p* = .20). The null hypothesis is retained.

Hypothesis 4a. Hypothesis 4a stated that participants in the FOR training conditions would have higher self-reported self-efficacy scores than the no-training condition. A Mann-Whitney U test revealed no significant difference in self-efficacy between the FOR condition (*Mdn* = 4.30) and the no-training condition (*Mdn* = 4.00), $U = 165.50$, $z = 1.43$, $p = .16$. Results indicated a small effect size $d = .25$ (Cohen, 1988). The small sample size could be the reason for the result, as the effect size suggest that there may be a significant effect in reality (Rosenthal & Rosnow, 2008). The null hypothesis is retained.

Hypothesis 4b. Hypothesis 4b stated that participants who scored higher self-reported self-efficacy scores would score higher accuracy scores. A Kendall's Tau test indicated that there was a significant positive relationship between the self-efficacy and rater accuracy ($r_\tau = .38$, $p = .00$). Therefore, the null hypothesis is rejected. Tables 5 and 6 below contain a summary for all the Kendall's Tau correlation and Mann-Whitney U t-tests.

Additional analyses

Further analyses were conducted to explore the aforementioned results. For example, personality and intelligence did not serve directly as a variable in the present study. The researcher wanted, however, firstly to analyse whether the variables were similar in both training conditions and secondly, whether there was any significant relationship between the two variables and the dependant variable, rater accuracy.

Intelligence. In terms of the differences between the two conditions, results revealed that participants in the FOR training condition (*Mdn* = 6, *SD* = 3.19) had slightly higher intelligence scores than the no-training condition (*Mdn* = 3, *SD* = 2.58). Whilst there appears to be a difference between the two conditions, a Mann Whitney test revealed there was no significant difference between the two groups ($U = .178.50$, $z = 1.92$, $p = .56$). Furthermore, a correlation test, using Kendall's Tau, revealed that whilst there was a small positive

relationship between accuracy and intelligence (r = .19), the relationship was insignificant (p = .15).

Personality. In terms of personality, no significant differences were found between the two training conditions: Extraversion (U = .118, z = -.379, p = .72), Agreeableness (U = .92, z = -1.37, p = .18), Conscientiousness (U = .118.50, z = -.359, p = .72), Emotional Stability (U = .147.50, z = .737, p = .47), and Openness (U = .167.50, z = 1.497, p = .14). Therefore, one can assume that both the FOR training condition and the no-training condition had no significant differences in personality. The Kendalls' Tau correlation test revealed that none of the Big Five personality traits significantly correlated with accuracy: Extraversion (r = -.07, p = .61), Agreeableness (r = -.09, p = .47), Conscientiousness (r = .21, p = .11), Emotional Stability (r = .11, p = .37), and Openness (r = .14, p = .27).

Manipulation checks. For the trainer manipulation check, both trainers were rated similarly in terms of enthusiasm (FOR: M = 4.56, SD = .54; No-training: M = 4.57, SD = .70; U = .108, z = -.93, p = .35) and differed only slightly in terms of professionalism (FOR: M = 4.18, SD = .57; No-training: M = 3.88, SD = .94; U = .102, z = -1.18, p = .24). This indicates the delivery of the training from the different trainers was fairly consistent across conditions and no significant differences were found between the training conditions. For the training manipulation check, participants in the FOR training condition reported a slightly higher score of their understanding of what the performance dimension was (FOR: M = 4.00, SD = .80; No-training: M = 3.94, SD = .93) and what to look for in the interviews following the training (FOR: M = 4.01, SD = .78; No-training: M = 3.88, SD = .86). A Mann Whitney U test revealed that there was no significant difference between the two training conditions in terms of knowing what the performance dimension was (U = .135, z = .28, p = .81) and knowing what to look for in the interviews following the training (U = .163.5, z = 1.46, p = .18).

Statistical power. Non-parametric tests have lower statistical power when compared to parametric tests, most notably with a small sample size (Corder & Foreman, 2009). This suggests that the study's results may have been affected by a lack of power (Aberson, 2010). To determine whether the null hypotheses have been incorrectly retained, a post-hoc power analysis was conducted to using G*Power. A benchmark statistic of .80 was used to indicate high statistical power (Cohen, 1988). Results from the post-hoc analysis revealed low statistical power for all tests conducted in the present study.

Table 5

Summary of Results: Kendall's Tau Correlation Test of Study Variables

Variable	1	2	3	4	Power
1. Accuracy	-				
2. Dispositional Reasoning	.17	-			.15[a]
3. Motivation	.17	-.07	-		.15[b]
4. Self-Efficacy	.38*	-.30	.60*	-	.59[c]

Note.

[a] = power statistic for correlation between accuracy and dispositional reasoning.

[b] = power statistic for correlation between accuracy and motivation.

[c] = power statistic for correlation between accuracy and self-efficacy.

*$p < .05$ (two-tailed).

Table 6

Summary of Results: Comparison of Training Groups using Mann-Whitney U Test

	FOR		NT						
Variable	*Mdn*	*SD*	*Mdn*	*SD*	U	*z*	*p*	*d*	Power
Accuracy	.70	.32	.18	.39	223	3.58	.00*	.68	.44
Dispositional Reasoning	15	1.46	13.50	2.66	193.50	2.49	.01*	.44	.22
Motivation	4.60	.23	4.57	.49	144	.61	.56	.11	.06
Self-Efficacy	4.20	.36	4.00	.54	165.50	1.43	.16	.25	.10

Note. FOR = frame-of-reference training condition; NT = no-training; *d* = Cohen's effect size.

*$p < .05$ (two tailed).

Discussion

The purpose of this study was to investigate how the well-known FOR training approach increases accuracy amongst raters. The researcher aimed to replicate previous findings (Powell & Bourdage, 2016; Powell & Goffin, 2009) to determine whether dispositional reasoning would influence the effectiveness of the FOR training approach in enhancing rater accuracy. The researcher also aimed to determine whether rater motivation and rater self-efficacy may influence the effectiveness of FOR training in enhancing rater accuracy, as motivation and self-efficacy remain relatively unexplored in contemporary research (De Kock et al., 2018).

The discussion is divided into five key sections. The first section discusses the main findings and contextualises the findings in relation to existing literature. The second section presents the theoretical implications this present study has added in relation to the knowledge surrounding rater accuracy and FOR training. The third section stipulates recommendation for future research. The fourth section addresses the limitations of the study. In the fifth section, the practical implications of the study are described in order for practitioners to understand and consider the results of the study. Finally, a summary of the findings is presented as a conclusion to this study.

Main Findings

In line with previous research (Roch et al., 2012; Woehr & Huffcut, 1994), it was expected that FOR training would increase rater accuracy. The results of this present study suggested that there was a notable difference in mean accuracy scores between those who received the FOR training and those who received no training. The large difference in accuracy scores between the two training conditions provides further support to the claim that FOR training can enhance accuracy (Roch et al., 2012; Woehr & Huffcut, 1994). In their

meta-analysis of rater training approaches, Roch et al. (2012) reported that FOR training held an average medium effect size.

Following the confirmation that FOR training increased rater accuracy in this present study, the researcher investigated how the training efforts affected individual difference constructs related to accuracy. The first individual characteristic speculated to be influenced by the training approach was dispositional reasoning. Dispositional reasoning is an ability that allows individuals to observe the behaviours of others and proceeds to make inferences about their personality traits, by taking into account situations, behaviours and co-occurrence of traits (Christiansen et al., 2005; De Kock et al., 2015). The researcher found that FOR training had a positive effect in enhancing dispositional reasoning, which suggests that dispositional reasoning can be developed.

What makes this finding interesting is that it is not in accordance with previous studies, who found FOR training attempts had no effect on dispositional reasoning (see Powell & Bourdage, 2016; Powell & Goffin, 2009). Therefore, previous findings suggest that the increase in accuracy in FOR training is not due to dispositional reasoning, as there was no difference between those who were trained through FOR and those who were not. The present study's findings however suggest that it is plausible that dispositional reasoning is responsive to training, and therefore may account for the difference in accuracy scores between those trained through FOR and those who were not.

A possible explanation for the contradiction between the present study and previous research in dispositional reasoning findings may be that the training intervention used in the present study was longer in duration, specifically 80 minutes, whereas the previous studies were approximately 30 to 60 minutes. This speculation is based on the argument that dispositional reasoning adheres to the classic criteria to be considered an intelligence rather than an innate ability (De Kock et al., 2018). Research has shown that intelligence may be

trained to a certain degree, however it is dependent on the amount of training received (Jaeggi, Buschkuehl, Jonides & Perrig, 2008). Therefore, it is credible to suggest that dispositional reasoning, as a type of intelligence, is responsive to FOR training, which may account for the increase in accuracy. Whilst this finding is tentative, it should be viewed positively as it sheds further insight into the ability-nature of dispositional reasoning (De Kock et al., 2015).

The second individual characteristic that was assumed to be influenced by the FOR training approach was rater motivation. Rater motivation refers to the rater's motivation to engage in the rating process and to provide accurate ratings (Ispas, 2010). According to Lievens (2001), training attempts would create 'buy-in' and would encourage raters to motivate accurately. The present study's findings revealed no significant difference in motivation scores between the raters who participated in the FOR training and those who were not.

A possible explanation for the lack of influence of FOR training on motivation is that perhaps both training groups were motivated. To elaborate, the no-training condition served as the control group, however participants were still involved in the workshop as well as the practice rating and discussion session. The only difference in the two training conditions is that the FOR training conditions received information and examples of the performance dimension, and the no-training condition did not. This may have resulted in the participants in the no-training condition assuming that they were involved in training. This could explain why there were no differences in motivation scores across the two training conditions (Lievens, 2001).

The third individual characteristic speculated to be influenced by the training approach was rater self-efficacy. Rater self-efficacy refers to a rater's belief and confidence in their ability to execute the behavioural demands of the rater role (Bernardin & Buckley,

1981). Although a notable difference existed in self-efficacy scores between those who received training and those who did not, however, this difference was not statistically significant. This could imply that self-efficacy may not be susceptible to be developed or influenced by training. This contradicts previous research which found that training raters to become more confident in the rating process was indeed possible (Bernadin & Villanova, 2005). An explanation for this contradiction may be that the training used by these previous researchers were specific self-efficacy focused training programmes, otherwise known as Self-Efficacy Training for Raters (SET-R). Perhaps, SET-R is able to develop self-efficacy but not the FOR training approach used in this study? This question should be explored in further research.

The final focus of the present study was to investigate whether the individual characteristics discussed above were able to predict rater accuracy. Rater self-efficacy was the only individual characteristic that the study found to be positively associated with rater accuracy. The effectiveness of rater training is largely dependent on the rater's belief that they can cope with difficulties and problems in the rating process by using their skills and knowledge gained in the training (Lievens, 2001; Wood & Marshall, 2008). If raters have this belief, then it is argued that they would be able to make more informed and accurate decisions in the rating process (Bernardin & Villanova, 2005).

The present study's results indicated that there was a significant positive relationship between the raters' self-efficacy scores and the accuracy of their ratings. Therefore, it is credible to posit that raters require the belief and confidence in their rating ability in order to make accurate ratings (Bernardin & Villanova, 2005). These findings replicate previous studies (Bernardin & Villanova, 2005; Wood & Marshall; 2008) who also found that self-efficacy was positively related to rater accuracy.

Previous researchers found that dispositional reasoning was found to be a moderate-to-strong predictor of rater accuracy (Christiansen et al., 2005; De Kock et al., 2015). However, no significant association between dispositional reasoning and rater accuracy was found in the present study. Whilst these findings contradict the assumption that dispositional reasoning leads to accurate ratings, it mirrors a previous study conducted by Powell and Goffin (2009), who found that although rater accuracy increased in their study following the training, they found no association between rater accuracy and dispositional reasoning. This lack of significance in our study may be due to either our small sample size, or alternatively, the possibility that studies which found a significant relationship between rater accuracy and dispositional reasoning focused on enhancing personality judgment accuracy (e.g. Christiansen et al., 2005; De Kock et al., 2015; Powell & Bourdage, 2016), whereas the present study focused on enhancing interview performance dimension rating accuracy.

In terms of rater motivation predicting rater accuracy, according to the social cognitive theory (Fiske & Taylor, 2013), it is believed that highly motivated raters will be more attentive and deliberately utilise judgement processes that focus on normative and systematic rating practises which are more accurate (Fiske & Taylor, 2013). The researcher, therefore assumed that motivation would positively influence accuracy, however results indicated that motivation had no association with rater accuracy.

A possible explanation for the lack of influence of rater motivation may rest with Harris (1994), who suggested that there are three determinant factors that affect rater motivation and its link to rater accuracy, namely: perceived rewards, perceived negative consequences and impression management concerns. To elaborate, it is believed that raters will be motivated to rate accurately should they perceive their accurate ratings will lead to receiving extrinsic rewards (Kanfer, 1990). A study showed that offering incentives to participants to rate accurately led to higher accuracy ratings (Salvemini et al., 1993). In this

present study, there were no perceived rewards to rate accurately in either training conditions. Due to ethical considerations, the researcher was not able to offer the FOR training an incentive to make accurate ratings and not the no-training condition, which would have affected the results.

The second determinant factor, perceived negative consequences, relates to the belief that a rater's motivation to making accurate ratings is influenced when they believe their ratings, accurate or inaccurate, could affect the relationship with colleagues (Harris, 1994; Murphy & Cleveland, 1991). For example, should a rater believe that inaccurate ratings would lead to negative consequences, such as an error in the selection of a candidate, then they would be more motivated to provide accurate ratings. In this present study there were no perceived negative consequences as participants were instructed to rate targets, not to make any decision based on these ratings. This eliminated any perceived negative consequences.

The last factor, impression management, influences motivation in the sense that the rater wishes to maintain an appropriate image to those involved in the rating context, such as the ratee or supervisor, and will adjust the accuracy of their ratings accordingly (Murphy & Cleveland, 1991). Impression management had no influence in this present study as the participants were not required to relay feedback to either the target being rated or the trainer. Therefore, participants were not compelled to consider or manage the effect of their ratings to others. Since none of these three factors were present in the present study, its findings support Harris's (1994) argument that motivation and its effect on rater accuracy is largely influenced by these factors.

Finally, the findings in this present study need to be considered in light of Funder's RAM framework model (1995; 1999; 2012). The researcher argued that rater dispositional reasoning, motivation and self-efficacy would increase the rater's cue detection and cue utilisation abilities. In other words, it was posited that these three individual characteristics

would increase a rater's ability to detect behavioural cues, relating to people management dimension, and to correctly utilise these cues when assigning ratings to each recorded interview. As previously mentioned, only self-efficacy was found to influence rater accuracy. This suggests that it is the only variable that may have affected the accuracy judgement process. This is credible as self-efficacy encourages raters to overcome difficulties in the rating process (Lievens, 2001), thereby allowing them to detect cues and utilise these cues more effectively.

In sum, the present study showed that FOR rater training may increase rater accuracy in a simulated interview context. In terms of the causal effect of FOR training on the individual characteristics being investigated in this study, dispositional reasoning was the only characteristic to be influenced by the training efforts. Furthermore, the study suggested self-efficacy may have played a role in increasing rater accuracy. However, these findings should be viewed as tentative.

Implications for Theory

This present study has offered new insight into FOR training and its established association with increased rater accuracy. The main contribution pertained to investigating the effectiveness of FOR training on a closer magnitude. The researcher did not simply assume that it was only due to rater ability or the training content that FOR training has an impact on rater accuracy. Prior research over the years has extensively confirmed that FOR training has a positive effect on the accuracy of ratings in the various Human Resource Management contexts (HRM; See Roch et al., 2012; Woehr & Huffcut, 1994). But few studies have explored how FOR training is effective (e.g. Dierdorff et al., 2010; Hauenstein & McCusker, 2018). There appeared to be a need for further research to be conducted to expand on why FOR training increases accuracy (Dierdorff et al., 2010) and the present study addressed this shortcoming.

In addition, the researcher further contributed to the understanding of dispositional reasoning, a variable which has been investigated through few empirical studies (De Kock et al., 2015; De Kock et al., 2018; Powell & Bourdage, 2016; Powell & Goffin, 2009). The main contribution was that the study revealed a difference in dispositional reasoning scores between the FOR training and no-training condition, which suggests that perhaps dispositional reasoning can be developed. Through this finding, the present study was able to support a previous claim made by De Kock et al. (2015) that dispositional reasoning may be an intelligence which can be developed (Jaeggi et al., 2008).

Furthermore, the study provided new insight into rater motivation and rater self-efficacy. This addressed the call by previous researchers to explore these relatively unexplored variables (De Kock et al., 2018; Spence & Keeping, 2010). Moreover, the study was able to show that rater self-efficacy may positively influence rater accuracy in a laboratory context. Further replication of these findings in field studies are recommended.

Lastly, a corollary of this study relates to generalisability issues. FOR training is traditionally applied in the performance appraisal context, where it was originally developed (Bernardin & Buckley, 1981). The present study showed that the FOR training approach can also be effective in enhancing interview accuracy. These findings mirrored previous researchers who successfully applied the FOR training approach in the employment interview context, similar to this study (Hauenstein, Facteau, & Schmidt 1999; Melchers, Lienhardt, von Aarburg, & Kleinmann, 2011). Therefore, one can assume that the FOR training approach is applicable to various HRM rater training contexts (Roch et al., 2012).

Future Research

Whilst the present study has provided further insight into the research area of how FOR training increases rater accuracy, the research area needs to grow in a few fruitful areas which deserve attention. Firstly, future research would benefit from replicating this study to

include the effect of FOR training on personality judgment accuracy as criterion (e.g., Christiansen et al., 2005; Powell & Bourdage, 2016; Powell & Goffin, 2009) in addition to interview performance dimension accuracy (e.g., De Kock et al., 2015) as the accuracy criterion.

Secondly, as the researcher found that rater motivation had no effect on the success of FOR training in increasing accuracy, further research needs to be conducted that focuses on the influence of three determinants factors to motivation, as proposed by Harris (1994). The researcher calls on further research to investigate whether these factors will enhance motivation during training and to study the consequential effect which motivation has on accuracy. As previously discussed, rater motivation has a promising theoretical underpinning that suggests motivation may influence the effectiveness of FOR training in increasing accuracy.

Lastly, future research needs to be conducted in order to further investigate how the individual characteristics in this study affect the accuracy judgement process as proposed by Funder's (1995; 1999; 2012) RAM framework. The present study did not measure how the individual characteristics specifically influence the accuracy process. Research would benefit from isolating the judgment accuracy stages, namely cue detection and cue utilisation, and measuring how each of the individual characteristics influence each of the stages.

Limitations

Although this present study offers new insight into rater accuracy and the FOR training approach, a few limitations need to be considered. First and foremost, this study was limited by a small sample size due to time and logistical constraints such as the availability and accessibility of students (Burns & Burns, 2008). Whilst this limitation was beyond the control of the researcher, it was foreseen. This limitation was addressed by offering incentive for participation as well as offering several sessions for students who differed in availability.

The small sample size limited the statistical analyses the researcher was able to conduct, specifically parametric statistics, which have been argued to be more rigorous than non-parametric tests (Field, 2013). Ideally, a MANOVA test would also have been desirable to assess the mediation effect of the study's variables on the dependent variable. A further limitation imposed by this smaller size, was the low statistical power. This placed a ceiling on the present study's attempts to achieve desired results (Cohen, 1988) and therefore limits the ability to conclude whether the effects of the variables investigated exists in reality (Field, 2013).

A second limitation of the present study was the generalizability of the study to the wider population of raters and judges in the various HRM rating contexts. Students were used from one course within a single university. The decision to use students was based on previous researchers defending the use of students (Christiansen et al., 2015; De Kock et al., 2015; Letzring, 2008; Powell & Goffin, 2009), as well as due to the time and logistical constraints in the present study. This has the potential of the sample group being too specific and raises questions about the degree to which results may generalise to the broader population of raters in organisations (Burns & Burns, 2008). Regardless, as study was experimental in design, the study able to control for potential confounding variables (Burns & Burns, 2008). By controlling variables, such as situational or participant variables, the present study was able to outweigh the limitation of conducting a laboratory study (Burns & Burns, 2008).

A third limitation is that only one performance dimension was rated. This may have simplified the rating context in the study, as most FOR training programmes have an average of five performance dimensions (Roch et al., 2012). This in turn may have over exaggerated the accuracy scores in the FOR training condition. Whilst this is a limitation, the decision to use one performance dimension was due to time constraints and possible effects of fatigue

(Cozby, 2007). FOR training requires a substantial amount of information to be relayed to the participants on each dimension and to provide as accurate frame-of-reference of each performance dimension as possible (Bernardin & Buckley, 1981). Future research would benefit from replicating this study and by implementing more performance dimensions to make the rating context more realistic (Roch et al., 2012).

Implications for Practice

One of the limitations in this study is viewed as a practical implication for research. Specifically, the small sample size is offered as a recommendation for future researchers to replicate the study on a larger scale. The results of this study are tentative and further research needs to be conducted in order to confirm our findings as to whether the FOR training approach is able to effectively enhance dispositional reasoning as well as to determine whether rater motivation and self-efficacy act as influencers on the effectiveness of FOR training in enhancing rater accuracy. In addition, further field studies similar to this present study are necessary for practice.

In terms of practical implications for organisations, perhaps judges and raters should be screened and selected based on their level of self-efficacy in their rating ability. The study implied that the more confident raters are, the more accurate they seem to be. This may result in more accurate ratings and consequentially more accurate and suitable selections. This would benefit organisations in multiple ways such as reduced staff turnover, increased productivity and profit margins (Bernardin & Villanova, 2005; Jiang et al., 2012).

Furthermore, research suggests that the more self-aware an individual is about their rating ability, the more they are able to develop their self-efficacy (Maddux & Gosselin, 2012). The development of self-efficacy is suggested to be influenced by the individual's ability to understand the cause and effect of their actions and self-reflection. Therefore, should organisations wish to further develop a rater's self-efficacy, organisations are

recommended to provide feedback to raters regarding their level of self-efficacy and the consequential effect on the accuracy of their ratings. This feedback would allow the rater to gain personal insight into their current level of self-efficacy. Personal insight should increase their self-efficacy (Maddux & Gosselin, 2012) which, in turn, may affect also their rating accuracy.

Conclusion

Rater accuracy is an important construct to the field of Human Resource Management (Christiansen et al., 2005). Training individuals to become more accurate raters is the forefront of rater accuracy research (Roch et al., 2012). This study investigated how the traditional FOR rater training method may lead to increased levels of rater accuracy. The purpose was to determine whether particular individual characteristics, namely rater dispositional reasoning, motivation and self-efficacy, would be influenced by the FOR training, and would consequentially also affect the accuracy of ratings. The study found that dispositional reasoning was responsive to the FOR training approach, which suggests that dispositional reasoning may account in part for the increase in accuracy following the FOR training. Although these individual characteristics have promising theoretical links to rater accuracy, the results of this study indicated that only self-efficacy predicted accuracy outcomes.

The present findings make a potentially valuable contribution to the understanding of how FOR training increases rater accuracy. Most importantly, this present study has lifted the veil on how rater characteristics play a part in the effectiveness of FOR training. The effect of these rater characteristics on rater accuracy in the FOR training approach should be further explored in field studies in order to provide insight as to how the FOR training enhances rater accuracy.

References

Aberson, C. L. (2010). *Applied power analysis for the behavioral science*. New York, NY: Psychology Press.

Allport, G. W. (1961). *Pattern and growth in personality.* Oxford: Holt, Reinhart & Winston.

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72,* 56-572. http://dx.doi.org/10.1037/0021-9010.72.4.567

Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive view*, Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Greenwich, CT: Information Age.

Banks, C. G., & Murphy, K. R. (1985). Toward narrowing the research-practice gap in performance appraisal. *Personnel Psychology, 38*, 335-345. http://dx.doi.org/10.1111/j.1744-6570.1985.tb00551.x

Benedict, M.E., & Levine, E.L. (1988). Delay and distortion: Tacit influences on performance appraisal effectiveness. *Journal of Applied Psychology, 73,* 507-514. http://dx.doi.org/10.1037/0021-9010.73.3.507

Bernardin, H. J., & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205-212. http://dx.doi.org/137.158.158.60

Bernardin, H. J., & Pence, E. C. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology, 65*, 60-66. http://dx.doi.org/10.1037/0021-9010.65.1.60

Bernardin, H. J., Tyler, C. L., & Villanova, P. (2009). Rating level and accuracy as a function of rater personality. *International Journal of Selection and Assessment, 17,* 300-310. http://dx.doi.org/10.1111/j.1468-2389.2009.00472.x

Bernardin, H. J., & Villanova, P. (2005). Research streams in rater self-efficacy. *Group & Organization Management, 30,* 61-88. http://dx.doi.org/10.1177/1059601104267675

Borman, W. C. (1977). Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance, 20,* 238-252. http://dx.doi.org/10.1016/0030-5073(77)90004-6

Bruner, J. S., & Tagiuri, R. (1954). The perception of people. In G. Lindzey (Ed.), *The handbook of social psychology* (Vol. 2, pp. 634-654). Reading, MA: Addison-Wesley.

Burns, R. P., & Burns, R. (2008). *Business research methods and statistics using SPSS*. London: Sage.

Campbell, D. T., & Stanley, J. C. (2015). *Experimental and quasi-experimental designs for research*. Chicago, IL: Ravenio Books.

Christiansen, N. D., Wolcott-Burnam, S., Janovics, J. E., Burns, G. N., & Quirk, S. W. (2005). The good judge revisited: Individual differences in the accuracy of personality judgments. *Human Performance, 18,* 123-149. http://dx.doi.org/10.1207/s15327043hup1802_2

Cleveland, J. N., & Murphy, K. R. (1992). Analyzing performance appraisal as goal-directed behavior. *Research in Personnel and Human Resources Management, 10,* 121-185.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates.

Condon, D. M., & Revelle, W. (2014). The International Cognitive Ability Resource: Development and initial validation of a public-domain measure. *Intelligence, 43,* 52-64. https://doi.org/10.1016/j.intell.2014.01.004

Cook, T. D., Campbell, D. T., & Day, A. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. Boston: Houghton Mifflin.

Corder, G. W., & Foreman, D. I. (2009). *Nonparametric statistics for non-statisticians: a step-by-step approach*. Hoboken, NJ: John Wiley & Sons.

Cozby, P. C. (2007). *Methods in behavioural research* (9th ed.). Fullerton: McGraw-Hill.

De Bruin, G. (2004). Problems with the factor analysis of items: Solutions based on item response theory and item parcelling. *SA Journal of Industrial Psychology, 30*, 16-26. http://hdl.handle.net/10520/EJC89027

De Kock, F. S., Lievens, F., & Born, M. Ph. (2015). An in-depth look at dispositional reasoning and interviewer accuracy. *Human Performance, 28*, 1-23. http://dx.doi.org/10.1080/08959285.2015.1021046

De Kock, F. S., Lievens, F., & Born, M. Ph. (2017). A closer look at the measurement of dispositional reasoning: Dimensionality and invariance across assessor groups. *International Journal of Selection and Assessment, 25*(3), 240-252, http://dx.doi.org/10.1111/ijsa.12176

De Kock, F. S., Lievens, F., & Born, M. Ph. (2018). The profile of the 'good judge' in HRM: A systematic review and agenda for future research. *Human Resource Management Review*. Advance online publication. https://doi.org/10.1016/j.hrmr.2018.09.003

Derous, E., Buijsrogge, A., Roulin, N., & Duyck, W. (2016). Why your stigma isn't hired: A dual-process framework of interview bias. *Human Resource Management Review, 26*(2), 90-111. http://dx.doi.org/10.1016/j.hrmr.2015.09.006

Dierdorff, E. C., Surface, E. A., & Brown, K. G. (2010). Frame-of-reference training effectiveness: Effects of goal orientation and self-efficacy on affective, cognitive, skill-based, and transfer outcomes. *Journal of Applied Psychology, 95*(6), 1181-1191. http://dx.doi.org/10.1037/a0020856

Djurdjevic, E. (2013). *The effects of social contextual factors on rater motivation and performance ratings* (Order No. 3603234). Available from ProQuest Dissertations & Theses A&I. (1469004676). Retrieved from https://search.proquest.com/docview/1469004676?accountid=14500

Driver, R. S. (1942). Training as a means of improving employee performance ratings. *Personnel, 18,* 364-370.

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia, PA: Psychology Press.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 33,* 56-70. http://dx.doi.org/10.1111/j.1745-3984.1996.tb00479.x

Eysenck, H. J. (1970). *The structure of human personality.* London: Methuen.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149-1160. http://dx.doi.org/10.3758/BRM.41.4.1149

Field, A. (2013). *Discovering statistics using IBM SPSS (4th ed.).* London: Sage.

Finkelstein, J. C. (1976). Experimenter expectancy effects. *Journal of Communication, 26*(3), 31-38. http://dx.doi.org/10.1111/j.1460-2466.1976.tb01900.x

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture*. London: Sage.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652-670. http://dx.doi.org/10.1037/0033-295X.102.4.652

Funder, D. C. (1999). *Personality judgment: A realistic approach to person perception.* San Diego: Academic Press.

Funder, D. C. (2012). Accurate personality judgment. *Current Directions in Psychological Science, 21,* 177-182. http://dx.doi.org/10.1177/0963721412445309

Funder, D. C., Furr, R. M., & Colvin, C. R. (2000). The Riverside Behavioral Q-sort: A tool for the description of social behavior. *Journal of Personality, 68*(3), 451-489. http://dx.doi.org/10.1111/1467-6494.00103

Funder, D. C., & West, S. G. (1993). Consensus, self-other agreement, and accuracy in personality judgment: An introduction. *Journal of Personality, 61*(4), 457-476. http://dx.doi.org/10.1111/j.1467-6494.1993.tb00778.x

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20-29. http://dx.doi.org/10.1111/j.1745-6916.2008.00058.x

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment, 4*, 26-42. http://dx.doi.org/10.1037/1040-3590.4.1.26

Graves, L. M. (1993). Sources of individual differences in interviewer effectiveness: A model and implications for future research. *Journal of Organizational Behavior, 14*(4), 349-370. http://dx.doi.org/10.1002/job.4030140406

Guion, R. M., & Highhouse, S. (2011). *Essentials of personnel assessment and selection.* Malwah, NJ.: Routledge.

Haaland, S., & Christiansen, N. D. (2002). Implications of trait-activation theory for evaluating the construct validity of assessment center ratings. *Personnel Psychology, 55*, 137-163. http://dx.doi.org/10.1111/j.1744-6570.2002.tb00106.x

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20*(4), 737-756. http://dx.doi.org/10.1016/0149-2063(94)90028

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93*(3), 258 -268. http://dx.doi.org/10.1037/0033-295X.93.3.258

Hauenstein, N. M. A., Facteau, J., & Schmidt, J. (1999). *Rater variability training: An alternative to rater error training and frame of reference training.* Paper presented at the 14th Annual Society of Industrial/Organizational Psychology, Atlanta, GA.

Hauenstein, N., & Foti, R. J. (1989). From laboratory to practice: Neglected issues in implementing frame-of-reference rater training. *Personnel Psychology, 42*(2), 359-378. http://dx.doi.org/10.1111/j.1744-6570.1989.tb00663.x

Hauenstein, N., & McCusker, M. E. (2017). Rater training: Understanding effects of training content, practice ratings, and feedback. *International Journal of Selection and Assessment, 25*(3), 253-266. http://dx.doi.org/10.1111/ijsa.12177

Hedge, J. W., & Kavanagh, M. J. (1988). Improving the accuracy of performance evaluations: Comparison of three methods of performance appraiser training. *Journal of Applied Psychology, 73*(1), 68. http://dx.doi.org/10.1037/0021-9010.73.1.68

Hedge, J. W., & Teachout, M. S. (2000). Exploring the concept of acceptability as a criterion for evaluating performance measures. *Group & Organization Management, 25*(1), 22-44. http://dx.doi.org/10.1177/1059601100251003

Huffcutt, A. I., Conway, J. M., Roth, P. L., & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology, 86*(5), 897-913. http://dx.doi.org/10.1037/0021-9010.86.5.897

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2013). Employment Interview Reliability: New meta-analytic estimates by structure and format. *International Journal of Selection and Assessment, 21*(3), 264-276. http://dx.doi.org/10.1111/ijsa.12036

Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational*

Behavior and Human Decision Processes, *54*(3), 321-368. http://dx.doi.org/10.1006/obhd.1993.1015

Ispas, D. (2010). The role of rater motivation in personnel selection validation studies. *Dissertation Abstracts International: Section B, Sciences and Engineering*, *71*, 7131.

Jackson, D. N., Chan, D. W., & Stricker, L. J. (1979). Implicit personality theory: Is it illusory? *Journal of Personality*, *47*, 1-10. http://dx.doi.org/10.1111/j.1467-6494.1979.tb00611.x

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences*, *105*(19), 6829-6833. http://dx.doi.org/10.3758/s13423-014-0699-x

Jiang, K., Lepak, D. P., Hu, J., & Baer, J. C. (2012). How does human resource management influence organizational outcomes? A meta-analytic investigation of mediating mechanisms. *Academy of Management Journal*, *55*(6), 1264-1294. http://dx.doi.org/10.5465/amj.2011.0088

Kanfer, R. (1990). Motivation theory and industrial and organizational psychology. In M. D. Dunnette, L. M. Hough, M. D. Dunnette, L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol. 1, 2nd ed* (pp. 75-170). Palo Alto, CA, US: Consulting Psychologists Press.

Kihlstrom, J.F., & Hastie, R. (1997). Mental representations of persons and personality. In S.R. Briggs, R. Hogan, & W.H. Jones (Eds.), *Handbook of personality psychology* (pp. 711-735). San Diego, Ca: Academic Press.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). Situational interview. *Journal of Applied Psychology*, *65*, 422-427. http://dx.doi.org/10.1037/t02627-000

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. *Journal of Applied Psychology, 60,* 550-555. http://dx.doi.org/10.1037/0021-9010.60.5.550

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment*, *6*(3), 141-152. http://dx.doi.org/10.1111/1468-2389.00085

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86*, 255-264. http://dx.doi.org/10.1037/0021-9010.86.2.255

Lievens, F., Schollaert, E., & Keen, G. (2015). The interplay of elicitation and evaluation of trait-expressive behavior: Evidence in assessment center exercises. *Journal of Applied Psychology*, *100*(4), 1169. http://dx.doi.org/10.1037/apl0000004

Letzring, T. D. (2008). The good judge of personality: Characteristics, behaviors, and observer accuracy. *Journal of Research in Personality, 42*(4), 914-932. https://doi.org/10.1016/j.jrp.2007.12.003

Levashina, J., Hartwell, C. J., Morgeson, F. P., & Campion, M. A. (2014). The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology*, *67*(1), 241-293. http://dx.doi.org/10.1111/peps.12052

London, M., Mone, E. M., & Scott, J. C. (2004). Performance management and assessment: Methods for improved rater accuracy and employee goal setting. *Human Resource Management, 43(4),* 319-336. http://dx.doi.org/10.1002/hrm.20027

Macan, T. (2009). The employment interview: A review of current studies and directions for future research. *Human Resource Management Review*, *19*(3), 203-218. http://dx.doi.org/10.1016/j.hrmr.2009.03.006

Maddux, J. E., & Gosselin, J. T. (2012). Self-efficacy. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd ed., pp. 198 –224). New York, NY: Guilford Press.

McIntyre, R. M., Smith, D., & Hassett, C. E. (1984). Accuracy of performance ratings as affected by rater training and perceived purpose of rating. *Journal of Applied Psychology, 69*, 147-156. http://dx.doi.org/10.1037/0021-9010.69.1.147

Melchers, K. G., Lienhardt, N., Von Aarburg, M., & Kleinmann, M. (2011). Is more structure really better? A comparison of frame-of-reference training and descriptively anchored rating scales to improve interviewers' rating quality. *Personnel Psychology, 64*(1), 53-87. http://dx.doi.org/10.1111/j.1744-6570.2010.01202.x

Mero, N. P., & Motowidlo, S. J. (1995). Effects of rater accountability on the accuracy and the favorability of performance ratings. *Journal of Applied Psychology, 80*, 517-524. http://dx.doi.org/10.1037/0021-9010.80.4.517

Millsap, R. E., & Maydeu-Olivares, A. (2009). *Handbook of quantitative methods in psychology.* Thousand Oaks, CA: Sage Publications Ltd.

Moser, K., Kemter, V., Wachsmann, K., Köver, N. Z., & Soucek, R. (2016). Evaluating rater training with double-pretest one-posttest designs: An analysis of testing effects and the moderating role of rater self-efficacy. *The International Journal of Human Resource Management*, 1-23. http://dx.doi.org/10.1080/09585192.2016.1254102

Murphy, K.R. & Cleveland, J.N. (1991). *Performance appraisal: An organizational perspective*. Boston, MA: Allyn & Bacon.

Nunnally, J. C., & Bernstein. I. H. (1978). *Psychometric theory*. New York: McGraw- Hill.

Powell, D. M., & Bourdage, J. S. (2016). The detection of personality traits in employment interviews: Can "good judges" be trained? *Personality and Individual Differences, 94*, 194-199. http://dx.doi.org/10.1016/j.paid.2016.01.009

Powell, D. M., & Goffin, R. D. (2009). Assessing personality in the employment interview: The impact of training on rater accuracy. *Human Performance, 22*, 450-465. http://dx.doi.org/10.1080/08959280903248450

Robinson, O. C. (2009). On the social malleability of traits: Variability and consistency in Big 5 trait expression across three interpersonal contexts. *Journal of Individual Differences, 30*, 201-208. http://dx.doi.org/10.1027/1614-0001.30.4.201

Roch, S. G. (2007). Why convene rater teams: An investigation of the benefits of anticipated discussion, consensus, and rater motivation. *Organizational Behavior and Human Decision Processes, 104*(1), 14-29. http://dx.doi.org/10.1016/j.obhdp.2006.08.003

Roch, S. G., McNall, L. A., & Caputo, P. M. (2011). Self-judgments of accuracy as indicators of performance evaluation quality: Should we believe them? *Journal of Business and Psychology, 26*(1), 41-55. http://dx.doi.org/10.1007/s10869-010-9173-6

Roch, S. G., Woehr, D. J., Mishra, V., & Kieszczynska, U. (2012). Rater training revisited: An updated meta-analytic review of frame-of-reference training. *Journal of Occupational and Organizational Psychology, 85*, 370-395. http://dx.doi.org/10.1111/j.2044-8325.2011.02045.x

Rosenthal, R., Kohn, P., Greenfield, P. M., & Carota, N. (1966). Data desirability, experimenter expectancy, and the results of psychological research. *Journal of Personality and Social Psychology, 3*, 20. http://dx.doi.org/10.1037/h0022604

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research* (3rd ed.). New York, NY: McGraw-Hill.

Salvemini, N. J., Reilly, R. R., & Smither, J. W. (1993). The influence of rater motivation on assimilation effects and accuracy in performance ratings. *Organizational Behavior and Human Decision Processes, 55*, 41-60. http://dx.doi.org/10.1006/obhd.1993.1023

Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*, 76-101. http://dx.doi.org/10.1006/obhd.1998.2751

Schmitt, N., & Chan, D. (1998). *Personnel selection: A theoretical approach*. Thousand Oaks, CA: Sage Publications.

Schneider, D. J. (1973). Implicit personality theory: A review. *Psychological Bulletin, 79,* 294-309. http://dx.doi.org/10.1037/h0034496

Sedikides, C., & Skowronski, J. J. (1991). The law of cognitive structure activation. *Psychological Inquiry*, *2*(2), 169-184. http://dx.doi.org/10.1207/s15327965pli0202_18

Smith, D. E. (1986). Training programs for performance appraisal: A review. *Academy of Management Review, 11*(1), 22-40. http://dx.doi.org/10.2307/258329

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. *Journal of Research in Personality, 68,* 69-81. https://dx.doi.org/10.1016/j.jrp.2017.02.004

Steers, R.M. & Porter, L.W. (1987). *Motivation and work behaviour.* New York: McGraw-Hill.

Spence, J. R., & Keeping, L. M. (2010). The impact of non-performance information on ratings of job performance: A policy-capturing approach. *Journal of Organizational Behavior*, *31*(4), 587-608. http://dx.doi.org/10.1002/job.648

Sulsky, L. M., & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497-506. http://dx.doi.org/10.1037/0021-9010.73.3.497

Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79*(4), 535-543. http://dx.doi.org/10.1037/0021-9010.79.4.535

Tett, R. P., & Burnett, D. D. (2003). A personality trait-based interactionist model of job performance. *Journal of Applied Psychology, 88*(3), 500-517. http://dx.doi.org/10.1037/0021-9010.88.3.500

Tett, R. P., & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34*, 397-423. http://dx.doi.org/10.1006/jrpe.2000.2292

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93*(3), 239-257. http://dx.doi.org/10.1037/0033-295X.93.3.239

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711-719. http://dx.doi.org/10.1037/0021-9010.93.3.711

Valentine, E. R. (1992). *Conceptual issues in psychology*. Boston: Allen & Unwin.

Woehr, D. J. (1994). Understanding frame-of-reference training: The impact of training on the recall of performance information. *Journal of Applied Psychology, 79*, 525-534. http://dx.doi.org/10.1037/0021-9010.79.4.525

Woehr, D. J., & Feldman, J. (1993). Processing objective and question order effects on the causal relation between memory and judgment in performance appraisal: the tip of the iceberg. *Journal of Applied Psychology, 78*(2), 232. http://dx.doi.org/10.1037/0021-9010.78.2.232

Woehr, D. J., & Huffcut, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205. http://dx.doi.org/10.1111/j.2044-8325.1994.tb00562.x

Wood, R.E., & Bandura, A. (1989). Social cognitive theory of organizational management. *Academy of Management Review, 14,* 361-384. http://dx.doi.org/10.2307/258173

Wood, R. E., & Marshall, V. (2008). Accuracy and effectiveness in appraisal outcomes: The influence of self-efficacy, personal factors and organisational variables. *Human Resource Management Journal, 18,* 295-313. http://dx.doi.org/10.1111/j.1748-8583.2008.00067.x

Yukl, G., Taber, T., Longenecker, C. O., Gioia, D. A., Sims, H. J., & Young, S. (1989). Power, politics, and performance appraisal. In J. W. Newstrom, K. Davis, J. W. Newstrom, K. Davis (Eds.), *Organizational behavior: Readings and exercises, 8th ed* (pp. 177-204). New York, NY, England: Mcgraw-Hill Book Company.

Appendix A: Ethics Clearance

Faculty of Commerce
Private Bag X3, Rondebosch, 7701
2.26 Leslie Commerce Building, Upper Campus
Tel: +27 (0) 21 650 4375/ 5748 Fax: +27 (0) 21 650 4369
E-mail: com-faculty@uct.ac.za
Internet: www.uct.ac.za
@Commerce_UCT **UCT Commerce Faculty Office**

11 July 2017

Ms Natasha Baret
School of Management Studies
University of Cape Town

REF: REC2017/07/005

Dear Ms Baret

Project: Why Does Frame-of-Reference Training Influence Rater Accuracy? A Test of Mediating Mechanism.

Thank you for submitting your study to the Faculty of Commerce Ethics in Research Committee.

It is a pleasure to inform you that the EiRC has **formally approved** the above-mentioned study.

Approval is granted for the period of 12 months. Should you require an extension or make any substantial changes to the research methodology which could affect the experiences of participants, you must submit a revised protocol to the Committee for approval.

Please note that the ongoing ethical conduct of the study remains the responsibility of the principal investigator.

Your sincerely

SAMANTHA ALEXANDER
Administrative Assistant
University of Cape Town
Commerce Faculty Office

Appendix B: Consent Form

Section of Organisational Psychology

School of Management Studies

Researcher: Natasha Baret

Study Aim:

We are interested in how you rate others in an interview.

Procedure:

You will participate in a training programme, after which you will complete six (6) questionnaires. One (1) of the six (6) questionnaires will be administered a week following the training via a link that will be distributed online. Feedback and results of the study will be made available through an academic dissertation and a potentially published journal article.

Incentive

We are offering three lucky draw research participation incentives of **R350 each**. **To qualify for these incentives, you will be required to complete the final questionnaire that will administered online a week following the training.**

Rights

This study has been approved by the Commerce Faculty's Ethics in Research (Approval Number: REC2017/07/005). Your participation is voluntary and your responses will be anonymous and used for research purposes only. You can choose to withdraw from the research at any stage. By signing below, you provide consent and agree that your answers will be used for research purposes.

I hereby give my consent by signing below:

Signature:_____

Please contact Natasha Baret if you have any questions regarding the research.